# ESTIMATING CONTEMPORARY MIGRATION RATES: EFFECT AND JOINT INFERENCE OF INBREEDING, NULL ALLELES AND MISTYPING

Juan J. Robledo-Arnuncio[1*] & Oscar E. Gaggiotti[2]

[1] *Department of Forest Ecology & Genetics, INIA-CIFOR, Ctra. de la Coruña km 7.5, 28040 Madrid, Spain.*

[2] *School of Biology, Scottish Oceans Institute, University of St Andrews, East Sands, St Andrews, Fife, KY16 8LB UK*

[*] *Corresponding author. E-mail: robledo.juan-jose@inia.es. Fax: +34 91 357 2293*

**Running headline:** Estimating plant migration rates

## Summary

1. Microsatellite-based genetic assignment is used broadly to monitor contemporary effective dispersal among populations. The need to investigate the robustness of this method to common genotyping errors was emphasized more than a decade ago, but it remains unaddressed.
2. We evaluate here for the first time the effect of mistaken and null alleles on estimates of contemporary seed and pollen migration rates obtained with genetic assignment methods. We also introduce a novel Bayesian approach to jointly estimate seed and pollen migration rates, genotyping error rates and null allele frequencies, not requiring independent reference or duplicate genotypic data.
3. Unaccounted-for mistaken alleles caused positive bias and increased the root mean square error (RMSE) of pollen migration rate estimates, whereas seed migration rate estimates were weakly sensitive to mistyping. Jointly estimating mistyping rates minimized the bias and RMSE they introduce on pollen migration estimates, while yielding seed migration rate estimates with similar or slightly larger bias and RMSE than those obtained when ignoring mistyping.
4. Ignoring genotyping errors can be especially problematic when there is no actual migration, because it can lead to the wrong conclusion that there is statistically significant exchange of pollen and/or seeds among populations that are actually isolated.
5. Unaccounted-for null alleles are problematic when among-population pollen dispersal is present, leading to underestimation of pollen migration rates and overestimation of seed migration rates. Jointly estimating null allele frequencies minimized these two biases, reduced the RMSE of seed migration rate estimates and produced relatively small changes in the RMSE of pollen dispersal estimates.
6. *Synthesis*. Disregarding genotyping errors and null alleles can produce biased and less accurate estimates of the rates at which present-day plant populations are exchanging seed and pollen. An approach is proposed here to minimize the effect of genotyping problems on contemporary migration rate estimates, which should help avoiding erroneous migration inference, monitoring and management, especially when dealing with low migration rates and their associated uncertainty.

**Key-Words:** Bayesian inference, dispersal, gene flow, genetic assignment, genotyping errors, microsatellite, pollen and seeds

## Introduction

Quantifying contemporary plant dispersal beyond current population boundaries is central to the evaluation of non-equilibrium demographic and evolutionary processes in changing environments. The rates of contemporary pollen and seed migration among populations condition many important processes, including metapopulation dynamics, speed of range shifts, reproductive assurance during colonization or fragmentation, adaptive or maladaptive gene exchange across heterogeneous habitats, and the potential risk of invasion, hybridization or genetic introgression posed by artificially translocated species or populations. In what follows, we will refer to contemporary migration rate as the proportion of effective propagules (pollen or seed) within a given population that have dispersed from an external population during a recent reference dispersal period, resulting in successful reproduction and/or seedling establishment.

Available methods based on genetic parentage assignment can estimate the contemporary rates of effective seed and pollen immigration from unknown sources into a given small population or study plot, provided all candidate parents within the latter are exhaustively genotyped (Slavov *et al.* 2005; Burczyk *et al.* 2006; Goto *et al.* 2006; Moran & Clark 2011). A more recent approach relaxes the exhaustive sampling requirement of previous parentage-based methods, potentially allowing accurate estimation of contemporary seed *or* pollen migration rates among a set of sampled populations (Wang 2014), but it still relies on a sufficient proportion of all parent candidates within each sampled population being genotyped (over 20% in practice; Wang 2014). Sampling for accurate parentage-based inference of among-population migration may thus be unfeasible over broad scales, unless populations are not large. However, populations of many plant and animal species frequently comprise many thousands of individuals, in which case genetic assignment is the method of choice for estimating contemporary migration.

Rather than aiming at parental identification, genetic assignment methods trace the origin of individuals to one or more populations, based on likelihood differences of the target individual's genotype across all candidate source populations, which are estimated from allelic frequencies in random reference samples of individuals within populations (Manel, Gaggiotti & Waples 2005). Some assignment methods require pre-defined populations as reference for individual

assignments (Paetkau *et al.* 1995; Rannala & Mountain 1997; Cornuet *et al.* 1999), while others simultaneously delineate populations and assign individuals to the inferred populations (Pritchard, Stephens & Donnelly 2000; Dawson & Belkhir 2001; Anderson & Thompson 2002; Guillot *et al.* 2005; Corander *et al.* 2008; Durand *et al.* 2009). Because of its ecological and evolutionary interest, some extensions of the former class of methods have focussed explicitly on unbiased estimation of the rate of contemporary migration (Pella & Masuda 2001, 2006; Wilson & Rannala 2003; Gaggiotti *et al.* 2004; Faubet & Gaggiotti 2008), and on its dissection into seed- and pollen-mediated components (Robledo-Arnuncio 2012; Unger, Vendramin & Robledo-Arnuncio 2014). Numerical simulation studies have shown that genetic assignment methods can generally estimate contemporary migration rates accurately at relatively low sampling cost (e.g., about 100 individuals per population typed at 10 polymorphic microsatellite loci in many cases) as long as genetic differentiation among candidate source populations is sufficiently large ($F_{ST} \geq 0.05$; Faubet, Waples & Gaggiotti 2007; Robledo-Arnuncio 2012). Weaker population differentiation can greatly increase the bias and variance of migration rate estimates, especially those of pollen migration, but errors can be reduced by increasing reference baseline samples to improve allele frequency estimation accuracy (Robledo-Arnuncio 2012).

Genetic assignment methods are suitable for assessing contemporary migration rates in non-equilibrium scenarios, such as ongoing fragmentation or range shifts, because they do not assume migration-drift equilibrium among candidate source populations and can account for Hardy-Weinberg disequilibrium (e.g. inbreeding) within populations (Wilson & Rannala 2003; François, Ancelet & Guillot 2006). However, they assume that marker loci are unlinked (but see Falush, Stephens & Pritchard 2003) and that there are no genotyping errors. Meeting the linkage equilibrium assumption is generally feasible by using appropriately developed markers, but low to moderate genotyping error rates are the norm for most available genetic assays and marker types (Pompanon *et al.* 2005), not least for microsatellites, which remain the most popular markers for contemporary migration inference (Selkoe & Toonen 2006). The need to investigate the robustness to genotyping errors of migration estimates produced by genetic assignment methods was emphasized more than a decade ago (Pompanon *et al.* 2005; Manel *et al.* 2005) but, despite the widespread use of these methods, it remains to be addressed. Given that population membership is ascertained based on allele frequencies, genetic assignment of individuals to

populations should not be as sensitive to genotyping errors as parentage assignment, for which mistyping directly leads to false parentage exclusion if unaccounted for (Pompanon *et al.* 2005; Wang 2014). Still, mistyping of genotypes can lead to both incorrect allele identification in candidate immigrant genotypes and inaccurate estimates of baseline population allele frequencies, which might upwardly bias migration estimates (Pompanon *et al.* 2005). Increasing stochastic typing errors could indeed decrease the apparent genetic differentiation among candidate source populations, and it has been shown that weaker differentiation leads to positively biased migration rate estimates (Faubet *et al.* 2007; Robledo-Arnuncio 2012). This might be especially problematic if there is no actual migration, in which case positively biased migration estimates could mislead evolutionary inference, ecological monitoring and conservation management. Thus, evaluating the extent to which genotyping errors could affect migration rate estimates is essential.

For microsatellites, null alleles (real alleles that consistently fail to amplify during PCR) are an additional common problem (Pompanon *et al.* 2005), which can result in both false parentage exclusion (Dakin & Avise 2004) and incorrect assignment of individuals to populations (Carlsson 2008). The precise effect on migration rate estimates is unknown and difficult to predict. Their tendency to produce wrong population assignments might induce positively biased migration estimates in some circumstances, but null alleles also increase apparent genetic differentiation among populations (Chapuis & Estoup 2007), which could reduce migration estimation errors. Moreover, null alleles will produce false homozygous diploid genotypes in candidate immigrants, with potentially contrasting effects on seed versus pollen migration estimates. Finally, null alleles should be (but have not been) factored into models inferring contemporary migration rates that account explicitly for Hardy-Weinberg disequilibrium, because the apparent increase in homozygosity produced by undetected alleles (Chybicki & Burczyk 2009) might impact the joint estimation of population inbreeding coefficients and migration rates.

In this study, we introduce genotypic likelihoods and a Bayesian scheme to jointly infer seed and pollen migration rates, population allelic frequencies, genotyping error (mistaken allele) rates, population inbreeding coefficients, and null allele frequencies. Using computer simulations, we investigate the behavior of the model and the effects of genotyping errors and null alleles on

147 contemporary seed and pollen migration rate estimates. We address the following practical
148 questions: Can genotyping errors and/or null alleles compromise contemporary seed and pollen
149 migration inference? Can we jointly infer contemporary migration rates and the frequency of
150 genotyping errors and null alleles? How much can we gain in estimation accuracy from such joint
151 inference (entailing substantial increase in parameter space dimensionality), as opposed to simply
152 ignoring genotyping errors and null alleles? Should we discard specific loci with high error rates
153 and null allele frequencies for better contemporary migration inference?
154

## Materials and methods

### *Demographic model and genotyping assumptions*

Building on Robledo-Arnuncio (2012), we assume a diploid plant species with discrete pollen and seed dispersal episodes. We are interested in estimating the proportion of immigrants from each of $I$ discrete external source populations among a sample of $D$ offspring (seeds or seedlings) collected in a target recipient population *after* a reference contemporary dispersal episode. A set of ($Q_0$, $Q_1$, . . ., $Q_I$) adult individuals are sampled from every population *before* the reference dispersal episode (the zero subindex refers to the recipient population hereafter). We assume that all source populations are sampled and that migration rates are low enough that there is a negligible joint probability of a seed sired by an immigrant pollen grain being in turn dispersed among populations (*double migration* event). The offspring sample can then be categorized into $2I$+1 groups: offspring born to local mothers fertilized by immigrant pollen (pollen migration), with proportions $\mathbf{m}_p = (m_{p1}, m_{p2}, \ldots, m_{pI})$, offspring born to nonlocal mothers fertilized by pollen from the same nonlocal population (seed migration), with proportions $\mathbf{m}_s = (m_{s1}, m_{s2}, \ldots, m_{sI})$ and offspring born to two local parents (local dispersal), with proportion $m_0$. Note that $m_0 + \sum_{i=1}^{I}(m_{pi} + m_{si}) = 1$. For notational convenience, we denote $\mathbf{m}$ the vector of size $2I$+1 containing the proportions of all offspring categories, obtained by concatenating the values of vectors $\mathbf{m}_p$ and $\mathbf{m}_s$ and the local recruitment rate $m_0$.

Both the adult and offspring samples are genotyped at $L$ unlinked codominant loci, yielding the $\mathbf{X} = \{X_{dl}\}$ and $\mathbf{Y} = \{Y_{iql}\}$ vectors of offspring and adult multilocus genotypes, where $X_{dl}$ is the diploid genotype of offspring $d$ at locus $l$, and $Y_{iql}$ is the diploid genotype of adult $q$ from population $i$ at locus $l$. The unknown adult pre-dispersal population allelic frequencies are given by a matrix $\mathbf{p} = \{p_{ila}\}$ giving the frequency of allele $a$ at locus $l$ in population $i$. We further assume that each locus $l$ has one or several null alleles (i.e. non-amplifying alleles) with unknown cumulative frequency $p_{il\text{-}}$ in population $i$. The rates of missing data for locus $l$ resulting from causes other than the presence of two null alleles are denoted $\beta_l$ and $\beta_l^*$ for offspring and adult samples, respectively. Non-null missing data may result from problems such as low-quality DNA

producing failed amplification of both alleles at a locus, which may differentially affect young versus adult plant tissues, and can also vary across loci (Buchan *et al.* 2005). We also consider the possibility of genotyping errors with locus-specific error rate $\mu_l$. Denoting $k_l$ the total number of (non-null) alleles observed over all populations at locus $l$, we assume that any true non-null allele (including the two homologous alleles in a genotype) is independently mistaken for any of the other $k_l - 1$ alleles with equal probability $\mu_l/(k_l - 1)$, for adult and offspring genotypes alike (Sieberts, Wijsman & Thompson 2002; Wang 2004). Mistaken alleles may result from DNA contamination, lack of specific amplification, or human errors during sample manipulation and data handling (Pompanon *et al.* 2005). Our current model is not intended to account for allelic dropout (stochastic non-amplification of one of the two alleles at a heterozygous locus) and false alleles (allele-like amplification artefacts; Pompanon *et al.* 2005).

We allow departures from Hardy-Weinberg equilibrium by assuming population-specific inbreeding coefficients, separately for offspring, $\mathbf{F} = (F_0, F_1, \ldots, F_I)$, and adults, $\mathbf{F^*} = (F_0^*, F_1^*, \ldots, F_I^*)$. Inbreeding coefficients reflect potential departures from random-mating within populations during the reference contemporary dispersal episode (**F**) and/or during former mating seasons producing adult cohorts (**F***). Among other factors, non-equilibrium demography (such as range shifts), long generation times, and inbreeding depression may generate differences in levels of population inbreeding between adults and offspring.

*Genotypic likelihoods*

Given the above assumptions and the unknown proportion of offspring from different origins **m**, population allelic frequencies before dispersal **p**, population inbreeding coefficients **F**, non-null missing data rates **β** and typing error rates **μ**, we can write the probability of observing multilocus genotype $X_d$ in the offspring sample of the target recipient population as

$$\Pr(X_d \mid \mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}) = \sum_{i=1}^{I}\left[m_{pi}\,\Pr_{i0}(X_d \mid \mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}) + m_{si}\,\Pr_{ii}(X_d \mid \mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\right] + m_0\,\Pr_{00}(X_d \mid \mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}) \quad \text{eqn 1}$$

where $\Pr_{ij}(X_d \mid \mathbf{p}, \mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\mu})$ is the probability of observing multilocus genotype $X_d$ in an offspring born to a mother from population $j$ pollinated by a father from population $i$, which for the $L$ unlinked loci corresponds to

$$\Pr_{ij}(X_d \mid \mathbf{p}, \mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\mu}) = \prod_{l=1}^{L} \Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \beta_l, \mu_l). \quad \text{eqn 2}$$

Consider first the case of null alleles without genotyping errors ($\mu_l = 0$ for all $l$). Probabilities of single-locus genotypes ($X_{dl}$) can be written depending on whether the male and female gametic phases originated from the same ($i = j$) or different populations ($i \neq j$), on whether any allele is observed at locus $l$ or not (i.e. missing data) and on whether the two eventually observed homologous alleles (denoted $a_1$ and $a_2$) are equal ($X_{dl}$ homozygous) or not ($X_{dl}$ heterozygous) (see Kalinowski & Taper 2006 and Chybicki & Burczyk 2009 for cases with $i = j$):

$$\Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \beta_l) = \begin{cases} (1-\beta_l)\left[(1-F_i)(p_{ila_1}^2 + 2p_{ila_1}p_{il-}) + F_i p_{ila_1}\right] & \text{if } i = j \text{ and } a_1 = a_2 \\ (1-\beta_l)\left[2(1-F_i)p_{ila_1}p_{ila_2}\right] & \text{if } i = j \text{ and } a_1 \neq a_2 \\ (1-\beta_l)\left[p_{ila_1}(p_{jla_1} + p_{jl-}) + p_{il-}p_{jla_1}\right] & \text{if } i \neq j \text{ and } a_1 = a_2 \\ (1-\beta_l)(p_{ila_1}p_{jla_2} + p_{ila_2}p_{jla_1}) & \text{if } i \neq j \text{ and } a_1 \neq a_2 \\ \beta_l + (1-\beta_l)\left[(1-F_i)p_{il-}^2 + F_i p_{il-}\right] & \text{if } i = j \text{ and missing data} \\ \beta_l + (1-\beta_l)p_{il-}p_{jl-} & \text{if } i \neq j \text{ and missing data} \end{cases} \quad \text{eqn 3}$$

With typing errors, the probability of observing (scoring) allele $a$ at locus $l$ in a randomly sampled gene from population $i$ is not $p_{ila}$, but rather $(1 - \mu_l)\, p_{ila} + \mu_l (1 - p_{ila} - p_{il-})/(k_l - 1)$, which must be factored into diploid genotypic probabilities. The resulting offspring genotypic likelihoods $\Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \beta_l, \mu_l)$ do not have such a simple form as equation 3 and are derived in equations S1 and S2 (see Appendix S1 in Supporting Information). The likelihood for the full set of $D$ offspring multilocus genotypes is then:

$$\Pr(\mathbf{X} \mid \mathbf{m}, \mathbf{p}, \mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\mu}) = \prod_{d=1}^{D} \Pr(X_d \mid \mathbf{m}, \mathbf{p}, \mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\mu}) \quad \text{eqn 4}$$

In order to estimate the unknown adult population allelic frequencies **p**, it is also necessary to formulate the likelihood of observed adult genotypes **Y** given **p**, **F***, **β*** and **μ**:

$$\Pr(\mathbf{Y} \mid \mathbf{p}, \mathbf{F}^*, \boldsymbol{\beta}^*, \boldsymbol{\mu}) = \prod_{i=1}^{I} \prod_{l=1}^{L} \Pr(\mathbf{Y}_{il} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l) \qquad \text{eqn 5}$$

where $\Pr(\mathbf{Y}_{il} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l)$ is the joint likelihood of the sample of $Q_i$ adult genotypes observed at locus *l* in population *i*. In absence of null alleles, typing errors and inbreeding, it would suffice to compute $\Pr(\mathbf{Y}_{il} \mid \mathbf{p}_{il})$ for each locus and population as in previous models (Rannala & Mountain 1997; Gaggiotti *et al.* 2004), using the vector of adult allelic counts $\mathbf{n}_{il} = \{n_{il1}, n_{il2}, \ldots, n_{ilk_l}\}$ and assuming a multinomial distribution $(\mathbf{n}_{il} \mid \mathbf{p}_{il}) \sim \text{Mult}(\mathbf{n}_{il}, \mathbf{p}_{il})$, where $n_{ila}$ is the observed number of copies of allele *a* at locus *l* among the $\mathbf{Y}_{il}$ genotypes sampled from population *i* at locus *l*, and the $\mathbf{p}_{il}$ are the adult population frequencies at locus *l* in population *i*. The multinomial distribution assumption for adult allelic counts holds if there are typing errors (following our genotyping error model) but neither null alleles nor inbreeding, as long as the probability vector $\mathbf{p}_{il}$ is replaced by $\mathbf{p}'_{il}$, with elements $p'_{ila} = (1 - \mu_l)\, p_{ila} + \mu_l (1 - p_{ila} - p_{il-})/(k_l - 1)$. If there are null alleles, however, allelic counts are compromised by the fact that it is not possible to determine a priori whether observed homozygotes carry two copies of the observed allele or a single copy plus a null allele. Adult inbreeding also violates the multinomial assumptions, as homologous alleles become non-independent. Null alleles and inbreeding, therefore, enforce the formulation of adult diploid genotypic likelihoods, rather than allelic counts, given not only **p** and **μ**, but also **F*** and **β***, which is the reason for introducing the two latter vectors in the model. Full expressions for $\Pr(\mathbf{Y}_{il} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l)$ are given in equation S3 (Appendix S2).

*Prior distributions of parameters*

We used uninformative prior distributions (*f*) for all parameters. The proportions of offspring from different origins (**m**) were assumed to follow a flat Dirichlet (symmetric with parameter equal to one) prior, $\mathbf{m} \sim \text{Dir}(\alpha=1)$. We also used a flat Dirichlet prior for population allelic

frequencies at each locus and population, $\mathbf{p}_{il} \sim \text{Dir}(\alpha=1)$. We assumed uniform priors on the interval (−1, 1) for offspring and adult population inbreeding coefficients of population $i$ ($F_i$ and $F_i^*$). The genotyping error rate at each locus ($\mu_l$) was assumed to follow a uniform prior on the interval (0, 1). Finally, we also assumed uniform priors on (0, 1) for offspring and adult non-null missing data rates at each locus ($\beta_l$ and $\beta_l^*$, respectively).

*Posterior distribution of parameters*

Given the **X** and **Y** vectors of genotypic data, the joint posterior distribution over parameter set $\boldsymbol{\Theta} = (\mathbf{m}, \mathbf{p}, \mathbf{F}, \mathbf{F^*}, \boldsymbol{\beta}, \boldsymbol{\beta^*}, \boldsymbol{\mu})$ is given by Bayes' rule:

$$f(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Y}) \propto \Pr(\mathbf{X} \mid \mathbf{m}, \mathbf{p}, \mathbf{F}, \boldsymbol{\beta}, \boldsymbol{\mu}) \Pr(\mathbf{Y} \mid \mathbf{p}, \mathbf{F^*}, \boldsymbol{\beta^*}, \boldsymbol{\mu})$$
$$f(\mathbf{m}) f(\mathbf{p}) f(\mathbf{F}) f(\mathbf{F^*}) f(\boldsymbol{\beta}) f(\boldsymbol{\beta^*}) f(\boldsymbol{\mu}), \qquad \text{eqn 6}$$

where the $f$ functions on the right-hand of the equation denote the prior distributions of individual parameters described above. We used the MCMC algorithm described in Appendix S3 to estimate the joint posterior distribution of equation 6.

*Simulation study of method performance*

Using the Monte Carlo algorithm detailed in Appendix S4, we investigated the expected bias, accuracy (root mean square error, RMSE) and credible interval non-coverage rate (NCR) of $\boldsymbol{\Theta}$ estimates obtained with equation 6 under different genotyping and migration assumptions, based on 250 stochastic replicates per combination of assumed parameter values. Our goal was to explore the potential effect of null alleles and mistyping on the estimation of contemporary seed and pollen migration rates, using neutral codominant molecular markers such as microsatellites. For each simulated scenario, we evaluated two alternative inferential approaches: using the *full model* that jointly estimates all parameters $\boldsymbol{\Theta}$ in equation 6 (including genotyping error rates and null allele frequencies), and using a *basic model* that only estimates **m**, **p** and **F**, ignoring mistyping and null alleles (i.e., implicitly assuming $\mu_l = 0$ and $p_{il-} = 0$ for every population $i$ and

298 locus *l*). Given that running the MCMC algorithm on the 250 simulated replicates for each
299 parameter combination was extremely time-consuming (more than 450 CPU hours), it was not
300 possible to explore a broad parameter space. For this reason, and because the sensitivity of pollen
301 and seed migration rate estimates to the number of external populations (*I*), population genetic
302 differentiation level ($F_{ST}$), and number of sampled offspring (*D*) and adults ($Q_i$) have been
303 already investigated in detail (Robledo-Arnuncio 2012), we fixed their values at $I = 2$, $F_{ST} = 0.1$,
304 $D = 100$, and $Q_i = 100$ (see, for example, Faubet *et al.* 2007 for a similar reference set). We also
305 fixed the rates of non-null missing data $\beta_l$ and $\beta_l^*$ at zero in the simulated data sets (although they
306 were assumed to be unknown and were jointly estimated), because preliminary analysis showed
307 that they were accurately estimated by the model and their effect on migration rate estimates was
308 similar to the effect of smaller sample sizes (*D* and $Q_i$), which has been already investigated
309 (Robledo-Arnuncio 2012). For the rest of parameters, we chose default reference values of $L = 10$
310 loci, $k_l = 10$ alleles/locus, population inbreeding $F = F^* = 0$, genotyping error $\mu_l = 0$, null allele
311 frequency $p_{\text{il}-} = 0$, and pollen and seed migration $m_{pi} = m_{si} = 0.05$; and considered the effect on
312 estimates of $\mathbf{m}$, $\mathbf{F}$, $\boldsymbol{\mu}$ and $\boldsymbol{p}_-$ of variable levels of $\mu$, $p_{\text{il}-}$, $F$ and $F^*$ (ranging from 0 – 20%), of $m_{pi}$
313 and $m_{si}$ (ranging from 0 – 5%) and of $L$ and $k_l$ (ranging from 5 to 20), varying only one or two
314 parameters at a time.
315

316 **Results**

317

318 No convergence problems in MCMC runs were observed for any of the simulated data sets,
319 which could be partly due to a choice of initial parameter values that could be recommended for
320 real data sets (Appendix S3). However, accurate estimation of offspring inbreeding coefficients
321 ($F_i$) of external populations was logically not possible in general, since the amount of information
322 for this purpose was either very small (when $m_{si} = 0.05$) or null (when $m_{si} = 0$). Although the
323 inference model reflected well this uncertainty in most runs, with rather uniform $F_i$ posterior
324 distributions, in some cases the MCMC chain got trapped at low negative values of $F_i$ of external
325 populations. This did not affect the estimation of other parameters. In particular, the estimation of
326 adult inbreeding coefficients (***F****) from parental genotypes was not problematic, while showing
327 very similar bias and RMSE than those of local offspring inbreeding estimates. For this reason,
328 we only report results for local $F_i$ when referring to inbreeding estimation below. We present here
329 results under different simulated scenarios, including (i) genotyping errors only, (ii) null alleles
330 only, (iii) null alleles and inbreeding, and (iv) genotyping errors, null alleles and inbreeding.
331 Errors in the estimation of immigration rates from the two separate external populations were
332 very similar in all scenarios considered (i.e. the model was able to discriminate immigration from
333 different sources), so we report average errors over the two external sources. Note that there is a
334 residual bias in seed and pollen migration estimates in all scenarios considered, even in total
335 absence of mistyping and null alleles. These biases could have been minimized by increasing
336 sample sizes, but we preferred considering values of the latter that are more typical in empirical
337 studies ($D = Q_i = 100$).

338

339 *Genotyping errors*

340

341 Using the basic inference model (ignoring genotyping error), the rate of mistaken alleles (μ) had
342 unequal effects on estimates of pollen ($\hat{m}_p$) and seed ($\hat{m}_s$) migration rates: $\hat{m}_p$ suffered
343 increasing bias and root mean square error (RMSE) with increasing μ, whereas $\hat{m}_s$ errors were
344 weakly influenced by μ (Fig. 1, white bars). Using the full model (jointly estimating genotyping
345 errors) resulted in the bias and RMSE of both $\hat{m}_p$ and $\hat{m}_s$ being rather insensitive to μ values

ranging from 0 to 20%, even if μ itself tended to be overestimated (Fig. 1, grey bars). The $\hat{m}_p$ bias reduction achieved by the full model was not accompanied by an increase in variance, and consequently the RMSE of $\hat{m}_p$ was lower for the full than for the basic model, the difference increasing with μ. By contrast, the bias and RSME of $\hat{m}_s$ were slightly larger for the full than for the basic model (Fig. 1). The credible intervals non-coverage rate (NCR) for both $\hat{m}_p$ and $\hat{m}_s$ were close to their nominal 5% value for both the full and basic models, even for high values of μ. Despite positively biased μ estimates, the NCR of μ remained below 10% for $\mu \leq 0.1$, increasing to about 20% for $\mu = 0.2$ (Fig. 1).

When there was no migration ($m_p = m_s = 0$), the benefits of accounting for genotyping errors were enhanced. Not only the positive bias of $\hat{m}_p$ was diminished when jointly estimating μ, but the RMSE of both $\hat{m}_p$ and $\hat{m}_s$ were lower for the full than for the basic model, independently of μ-value (Fig. 2). In addition, genotyping errors without migration resulted in higher than nominal NCR for both $\hat{m}_p$ and $\hat{m}_s$ when using the basic model, but not when using the full model (Fig. 2).

Assuming marker loci with variable mistyping rates (see Fig. 3), improvement in the estimation of $m_p$ or $m_s$ was neither achieved by discarding all loci with $\mu \geq 0.05$ (*strict* strategy), nor by discarding those with $\mu \geq 0.1$ (*medium* strategy), as compared to the *relaxed* strategy where no locus was discarded. The RMSE of $\hat{m}_s$ was very similar for all three data-filtering strategies, while that of $\hat{m}_p$ increased for the *medium* and especially for the *strict* strategy. These results held irrespective of whether error rates were jointly estimated or not (Fig. 3).

Varying simultaneously the number of loci ($L$) and alleles per locus ($k_l$), holding $L*k_l$ constant, did not change qualitatively most of the effects of μ on migration estimates observed for the default values $L = k_l = 10$ (Fig. 1 versus Figs. S1 and S2 in Supporting Information). More precisely, regardless of the assumed values of $L$ and $k_l$, we observed positive biases of $\hat{m}_p$ induced by genotyping error, which largely decreased when jointly estimating μ; weak sensitivity of $\hat{m}_s$ to genotyping error; lower RMSE of $\hat{m}_p$ obtained using the full versus the basic model,

374 especially for large μ; and slightly larger RMSE of $\hat{m}_s$ obtained using the full versus the basic
375 model, for most values of μ. However, in the case of the full model, reducing $k_l$ to five (and
376 increasing $L$ to 20) resulted in increased overestimation of genotyping error rates, in turn
377 producing slight underestimation of $m_p$ and increased overestimation of $\hat{m}_s$ (Fig. S1 versus Fig.
378 1). Increasing $k_l$ to 20 (and reducing $L$ to five), by contrast, resulted in genotyping error rates
379 being underestimated rather than overestimated, with limited impact on migration rates estimated
380 with the full model (Fig. S2 versus Fig. 1).
381
382 *Null alleles*
383
384 When using the basic inference model (which ignores null alleles), increasing null allele
385 frequencies ($p_-$) increased underestimation of $m_p$ and overestimation of $m_s$ (Fig. 4), which ensued
386 from the basic model wrongly perceiving some falsely homozygous pollen immigrants as seed
387 immigrants. Jointly estimating the frequency of null alleles with the full model largely eliminated
388 the increase in bias of both $\hat{m}_p$ and $\hat{m}_s$, with their RMSE being fairly insensitive to null alleles
389 (Fig. 4). For the basic model, an increase in migration rate biases translated into a larger RMSE
390 for $\hat{m}_s$ but not for $\hat{m}_p$ (except for $p_- = 0.2$), as the variance of $\hat{m}_p$ (but not of $\hat{m}_s$) was decreased
391 by the presence of null alleles. Overall, the RMSE of $\hat{m}_p$ was rather similar for the full and basic
392 models (except for $p_- = 0.2$, in which case the full model exhibit lower RMSE), while the RMSE
393 of $\hat{m}_s$ was lower for the full model, the difference increasing with increases in $p_-$. The NCR for
394 both $\hat{m}_p$ and $\hat{m}_s$ was close to or below the nominal value for the full model, whereas for the
395 basic model NCR was above nominal for $p_- \geq 0.1$, and above 40% in the case of $\hat{m}_p$ when $p_- =$
396 0.2. An additional effect of null alleles was the strong positive bias and RMSE of population
397 inbreeding ($F$) estimates obtained with the basic model (Fig. 4), which were greatly reduced
398 when jointly estimating null allele frequencies. The full model tended however to underestimate
399 $p_-$ and overestimate $F$ for high values of $p_-$ (Fig. 4).
400

The mechanism behind the opposite-sign biases of $\hat{m}_p$ and $\hat{m}_s$ induced by null alleles became evident when assuming that either seed or pollen migration was zero. If there is pollen but not seed migration ($m_p > 0$, $m_s = 0$), then null alleles had very similar qualitative and quantitative effects on parameter estimates as in the scenario with both seed and pollen migration (Fig. S3 versus Fig. 4), since null-allele-driven false homozygosity still results in the basic (but not the full) model wrongly perceiving some pollen immigrants as seed immigrants. By contrast, in absence of pollen migration ($m_p = 0$, $m_s > 0$), the bias increment of $\hat{m}_s$ produced by null alleles largely disappears, because $m_p$ cannot be underestimated (Fig. S4). In addition, if $m_p = 0$ then the bias and RMSE of both $\hat{m}_p$ and $\hat{m}_s$ are lower for the basic model than for the full model, irrespective of null allele frequency (Fig. S4).

In the simulated scenarios where assumed null allele frequencies varied across loci (see Fig. 5), discarding loci with $p_- \geq 0.1$ (*strict* strategy) or with $p_- \geq 0.2$ (*medium* strategy) worsened the RMSE of $\hat{m}_p$, irrespective of whether $p_-$ was jointly estimated or not, as compared to the *relaxed* strategy where no locus was discarded. This was the case even if the *medium* strategy reduced the bias of $\hat{m}_p$ estimated by the basic model. The bias of $\hat{m}_s$ was minimized adopting the *strict* strategy, for both the basic and full models, while its RMSE was insensitive to genotyping strategy for the full model and, in the case of the basic model, decreased only slightly when removing loci with null alleles (Fig. 5). NCR of migration estimates were close to or below nominal value in all cases.

*Population inbreeding and null alleles*

Assuming simultaneous presence of null alleles and population inbreeding did not alter the effects of null alleles on migration rate estimates observed under random mating ($F = 0$; Fig. S5 versus Fig. 4). Moreover, the full model still adequately discriminated $F$ and $p_-$, though again somewhat overestimating $F$ and underestimating $p_-$ (with their respective NCR exceeding nominal level) when their assumed values were as high as 0.2 (Fig. S5). On the other hand, when assuming population inbreeding but not null alleles, increasing $F$ values had minimal impact on

the bias and RMSE of $\hat{m}_p$ and $\hat{m}_s$, equally for the basic and full models (Fig. S6). The full model tended however to overestimate null allele frequencies when they are zero (although the NCR of $p_-$ exceeded nominal values only slightly), which caused underestimation of $F$ (Fig. S6).

*Genotyping errors, null alleles and population inbreeding*

We considered a reference simulated scenario with neither mistyping, nor null alleles or inbreeding (scenario *A*: $F = \mu = p_- = 0$), and compared it with three others with moderate rates of both mistyping and null alleles, two of which without inbreeding (scenario *B*: $F = 0$, $\mu = 0.05$, $p_- = 0.1$; and scenario *C*: $F = 0$, $\mu = 0.1$, $p_- = 0.1$) and one with inbreeding (scenario *D* with $F = \mu = p_- = 0.1$). When null alleles and mistyping were present but ignored (using the basic model), the effects of null alleles on migration estimates appeared to prevail over those of mistyping, namely $\hat{m}_p$ exhibited increased negative (rather than positive) bias and slightly reduced (rather than increased) variance and RMSE, while $\hat{m}_s$ showed increased bias and RMSE (scenarios *B*, *C* and *D*; Fig. 6). As compared to the basic model, jointly estimating all parameters with the full model reduced the bias of both $\hat{m}_p$ and $\hat{m}_s$ as well as the RMSE of $\hat{m}_s$ in all scenarios, while the RMSE of $\hat{m}_p$ was rather similar for the two models (slightly larger for the full model except in scenario *C*; Fig. 6). The NCR of $\hat{m}_p$ and $\hat{m}_s$ did not exceed the nominal value in any of the scenarios when estimated using the full model, while they did exceed it in scenarios *B*, *C* and *D* when using the basic model (Fig. 6).

## Discussion

We have for the first time evaluated the effect of genotyping errors and null alleles on estimates of contemporary migration obtained with genetic assignment methods. We have also introduced a novel Bayesian approach to estimate jointly seed and pollen migration rates, genotyping error rates and null allele frequencies, which does not require independent information on error rates derived from reference or duplicate data.

Microsatellite-based genetic assignment is used broadly to monitor contemporary effective dispersal among populations but such estimates can be biased due to common microsatellite mistyping and null alleles. Thus, our inference model and numerical results should be of interest to statistical and empirical ecologists dealing with plant population dynamics in non-equilibrium systems. Our results showed that, although not always serious, unaccounted for mistaken or null alleles may compromise accurate estimation of contemporary seed and/or pollen migration rates and their associated uncertainty. In many of the simulated scenarios considered, jointly inferring mistyping rates and null allele frequencies reduced migration estimation errors caused by these genotyping problems, notably when these errors were large. We discuss here first the main qualitative trends found over the entire parameter space considered in our simulation study, focusing then on practical consequences for typical empirical values of mistyping and null alleles.

### *Effect and joint inference of genotyping errors*

Under our demographic and genotyping assumptions, unaccounted-for mistaken alleles resulted in some local recruits being perceived by the model as pollen immigrants (by which we mean having greater model likelihood of being such), consequently positively biasing pollen migration estimates and increasing their RMSE. This result is consistent with predictions that the decrease in apparent genetic differentiation produced by mistyping should result in migration overestimation (Pompanon *et al.* 2005). Obviously, most mistyped local recruits should still be correctly perceived by the model as local, while stochastically mistaken alleles should also result

in some mistyped pollen immigrants being wrongly perceived by the model as locally recruited. Both effects, local recruits being perceived as pollen immigrants and pollen immigrants being perceived as local recruits, can certainly occur simultaneously (unless migration is null), but the latter should be less important than the former in absolute terms, because we realistically assumed lower among- than within-population pollination (had we assumed that the majority of local offspring were pollen immigrants, mistyping would then have tended to induce negatively biased migration estimates). The larger proportion of local versus migrant offspring would produce mistyping-induced systematic biases in pollen migration rate estimates despite the expectation that genotyping errors should not cause systematically biased population assignments (c.f. Wang 2014).

The identification of seed immigrants did not appear to be noticeably impaired by unaccounted-for mistyping. Seed migration rate estimates have been shown to be far more robust to demographic and sampling assumptions than pollen migration ones (Robledo-Arnuncio 2012), since seed immigrants carry twice as much allelic information about their biparental origin than pollen immigrants do about their paternal one. The same fact should render seed migration estimates less sensitive to mistyping. Nevertheless, our result that they were virtually insensitive even to very high levels of mistyping is noteworthy. The explanation for this observation is that seed genotypes with just one or a few loci with two (non-mistyped) homologous alleles that are *relatively* rare across candidates other than the true source may be enough to yield strongly discriminant genotypic likelihoods across candidate populations. This result held even under assumed levels of population differentiation as low as $F_{ST} = 0.01$ (results not shown).

Jointly estimating mistyping rates with our full model minimized the bias and RMSE increase that they caused on pollen migration estimates, while yielding seed migration estimates that had similar or slightly larger bias and RMSE than those obtained when ignoring mistyping. Mistyping rates themselves were overestimated when assuming ten and especially five alleles per locus, while they were underestimated when using loci with 20 alleles. The increasing overestimation of mistyping rates with decreasing numbers of alleles could be related with an identifiability problem in our genotyping error model (which should reflect reality to the extent that our model assumptions hold). Specifically, it can be shown that the assumption of allele-independent

genotyping errors determines that, if alleles are equifrequent within a population, then observed population allelic and genotypic frequencies are not influenced by randomly mistaken alleles, in which case the model cannot detect typing errors. The impact of mistyping on observed allelic and genotypic frequencies, and therefore the chance to infer mistyping rates, increases with increasingly skewed allelic frequency distributions, which were more frequent under the scenarios with the largest number of alleles. The problem would be greatest for biallelic markers such as SNPs, for which models explicitly formulated in terms of genotype-dependent mistyping rates (e.g. Kalinowski, Taper & Marshall 2007) should be more adequate, especially because SNP scoring typically involves genotype (rather than allele) calling through cluster analysis. Our assumed mistyping model should remain valid, however, to describe several common sources of stochastic error in microsatellite genotyping (Sieberts *et al.* 2002; Wang 2004).

Overall, in terms of accuracy (RMSE), our simulations suggest that the improvement in pollen migration estimates achieved by jointly estimating mistyping generally exceeded substantially the potential worsening in seed migration estimates (Figs. 1, 2, 3, S1 and S2). Simulations also showed that information loss exceeded noise reduction when discarding mistyping-prone loci, not improving, but rather worsening, migration estimates. The best strategy for accurate estimation of seed and pollen migration rates would thus seem to be using all available loci and jointly estimating mistyping rates. The special case of zero actual migration is noteworthy, because the full model not only produced less biased and more accurate estimates of seed and pollen migration rates in that case, but also avoided the inflated non-coverage rates of credible intervals for migration estimates obtained ignoring mistyping. This result suggests that ignoring mistyping could frequently lead researchers to the wrong conclusion that there is a statistically significant positive seed and/or pollen migration rate (i.e. that there is some demographic and genetic cohesion) among populations that are actually not exchanging any migrant whatsoever, which, besides being ecologically misleading (e.g. Waples & Gaggiotti 2006), could be specially problematic for conservation management (Ellstrand 1992; Mills & Allendorf 1996) and risk assessment (Laikre *et al.* 2010).

*Effect and joint inference of null alleles*

543 The effects of not accounting for null alleles differed between estimates of seed and pollen
544 migration rates. Considering any given locus of an offspring born to a local mother and an
545 external father (i.e., a pollen immigration event), if the paternal immigrant allele is null then the
546 observed genotype will be homozygous for the maternal local allele, leading to underestimates of
547 pollen migration and overestimates of local dispersal. Whereas if the maternal local allele is null,
548 then the observed genotype will be homozygous for the immigrant paternal allele, leading to
549 underestimates of pollen immigration and overestimates of seed immigration. Taken together, the
550 two possibilities should induce bias increments of opposite sign, negative for pollen migration
551 and positive and weaker for seed migration, which is what we observed in the simulations that
552 assumed there is actual pollen migration (with or without seed migration). By contrast, if any of
553 the two homologous alleles of a seed immigrant is null, its observed genotype would still display
554 two (equal) immigrant alleles, which should not bias migration estimates. The capability of null
555 alleles to mask pollen but not seed immigrants was reflected in the simulations with seed but no
556 pollen migration, in which no bias increase was observed with increasing null allele frequency,
557 neither for seed nor for pollen migration estimates.
558
559 Null alleles had the additional effect of reducing the variance of pollen (but not seed) migration
560 estimates, probably because of increased apparent population genetic differentiation (Chapuis &
561 Estoup 2007). This effect is consistent with the stronger sensitivity of pollen migration rate
562 estimates to actual population differentiation (Robledo-Arnuncio 2012), and was strong enough
563 to override the simultaneous detrimental bias increment caused by null alleles, resulting in
564 unchanged or more accurate (with lower RMSE) estimates of pollen dispersal at intermediate
565 (<20%) than at zero null allele frequencies. By contrast, the accuracy of seed migration estimates
566 was largely driven by their bias, both augmenting with increasing null allele frequencies except in
567 absence of pollen migration (in which case they were insensitive to nulls).
568
569 If there is pollen migration (with our without seed migration), jointly estimating the frequency of
570 null alleles minimized the bias increase that they caused in both seed and pollen migration
571 estimates. The bias correction translated into consistently lower RMSE for seed dispersal
572 estimates that accounted for, rather than ignored, null alleles, while it produced relatively smaller
573 changes in the RMSE of pollen dispersal estimates, which sometimes increased and sometimes

decreased (notably when the frequency of nulls was as high as 20%) relative to the basic model ignoring nulls (Figs. 4, S4 and S5). Altogether, jointly estimating null allele frequencies (without discarding any locus from the analysis) seems a better inferential strategy if pollen migration is present, because the improvement in the accuracy of seed migration estimates exceeds the potential worsening in that of pollen migration for the range of parameters considered. If pollen migration was actually zero, jointly estimating null allele frequencies had virtually no impact in seed migration estimates and it increased the RMSE of pollen migration estimates relative to the basic model ignoring null alleles, but always producing credible intervals with nominal coverage for both seed and pollen migration rate estimates. Therefore, if there is no prior knowledge on the level of pollen migration, accounting for null alleles seems, overall, advisable. If pollen migration were known to be zero a priori, however, researchers could then focus on the estimation of seed migration rates and set pollen migration rates at zero. In the latter case, accounting for null alleles would make little difference on seed migration (though they should be accounted for if population inbreeding coefficients are also of interest).

*Practical considerations*

Our model is relevant for a general class of study systems involving discrete plant populations: an offspring sample collected in a recipient population in which there are local reproductive adults (producing pollen and seeds), and that may be receiving a small proportion of pollen (and possibly seed) immigrants from external sources. This scenario may correspond, for example, to a rear-edge population, or to a front-edge population established some generations ago, for which we want to establish the level of present-day connectivity via seed and pollen dispersal with a set of external populations. The rates of pollen and seed immigration will be unknown a priori in this general case, although for many plant species we know that among-population migration takes place predominantly through pollen flow (Ennos 1994; Kremer *et al.* 2012), so migration, when present, will most likely involve either both pollen and seed migration, or only pollen migration. Our numerical analysis suggests that neither mistyping nor null alleles should be disregarded when inferring migration in this kind of scenario, especially if rigorous uncertainty assessment is required.

A different scenario of potential interest could be an offspring sample collected in a recipient population with no local reproductive adults, having originated via seed dispersal from external populations. This could correspond to an ongoing colonization, where the question at hand is the proportion of recently established non-reproductive individuals (e.g. saplings forming a stand in the front edge of a tree species distribution) that have originated via seed dispersal from different potential external sources. In this case, effective pollen immigration into the target stand would be known to be (and could be fixed at) zero a priori, and it would be safe to ignore null alleles for seed migration inference. Although in principle typing errors would also be of less concern in this scenario, because estimates of seed migration rates seem rather robust to moderate-to-high mistyping rates, they might somewhat distort the estimated distribution of migration rates from different sources if their contributions to the migrant seed pool is very unequal. In particular, unaccounted-for mistyping could produce spurious statistically significant migration estimates from candidate sources from which seeds have not migrated (Fig. 2), along with slightly underestimated migration rates from the actual sources. This problem would be alleviated by jointly estimating genotyping error rates.

Available studies suggest that per-locus genotyping error rates are generally below 10% (Bonin *et al.* 2004; Hoffman & Amos 2005; Morin *et al.* 2009), and null allele frequencies usually lower than 20% (Dakin & Avise 2004). Encouragingly, our results would suggest that errors in migration estimates caused by these levels of unaccounted-for mistaken and null alleles are not likely to be large in many cases, even under the assumption that all assayed loci are simultaneously affected by such levels of both typing problems (Fig. 6). Besides not being large, these errors may generally be reduced by jointly estimating migration rates, null allele frequencies and mistyping error rates with our model. This would additionally yield better population inbreeding estimates and credible intervals with better coverage. In any case, the effect of mistyping and null alleles on migration estimates will be influenced by the actual distribution of parental allelic frequencies within and among populations. Thus, it is highly advisable to conduct pilot numerical analyses (as suggested by Pompanon *et al.* 2005) to evaluate effects of different error rates and null allele frequencies on method performance for particular genetic assays, empirical reference baseline genotypes and assumed levels of migration. Of course, efforts should focus first on minimizing the production of errors during genotyping,

especially in the offspring sample, because incorrect allele identification in candidate immigrant genotypes induces greater errors in migration estimates than inaccurate estimates of baseline population allele frequencies (results not shown).

*Model limitations and perspectives*

Our model represents a first formal approach to inferring contemporary migration jointly with genotyping error rates and null allele frequencies. For the sake of tractability, we necessarily made several simplifying assumptions, as in previous error models (Sieberts *et al.* 2002; Wang 2004). Notably, we assumed that genotyping errors are allele-independent, that homologous alleles at a locus (irrespectively of whether they are equal or not) are independently mistaken, and that errors are independently distributed across loci. The consequences for migration inference of typing errors violating these assumptions are not easily predictable and could be the subject of future numerical studies, as well as the effect of other sources of error not considered here, such as allelic dropouts (see Wang 2004) and false alleles. Formally accommodating all major distinct sources of error of particular markers into genotypic likelihoods usable for migration inference will be challenging, and the extent to which different errors would be identifiable, and whether there is much to be gained by such complex statistical exercise, are unclear.

Other possible extensions of the model presented here could be the incorporation of available Bayesian schemes for jointly inferring the effect of environmental factors on migration (Gaggiotti *et al.* 2004; Faubet & Gaggiotti 2008), or for estimating migrant proportions with incomplete baseline samples (Dawson & Belkhir 2001; Pella & Masuda 2006), as unsampled populations could augment mistyping-induced positive biases in pollen migration estimates from sampled sources. Research efforts in the field are now heading to the combination of mechanistic and genetic approaches for better estimation of broad-scale seed and pollen migration rates and their main spatial, ecological and demographic determinants (Robledo-Arnuncio *et al.* 2014). Plant dispersal ecology advances will require models for the rate and paths of contemporary seed and pollen migration in spatially, environmentally and demographically explicit context, while properly accounting for different sources of uncertainty, especially for rare but important long-distance dispersal events. The approach proposed here contributes to this endeavor with a

667 framework to account for potential errors in contemporary migration estimates caused by
668 genotyping problems, which should prove especially useful for accurately assessing low levels of
669 migration and associated uncertainty.
670
671 A program written in C++ language implementing the estimation method described in this article
672 will soon be freely available at https://sites.google.com/site/espmsoftware or by contacting J. J.
673 R-A.
674

675 **Acknowledgements:** This work was supported by CGL2015-64164-R project from the Spanish
676 Ministry of Economy and Competitiveness and the European Regional Development Fund.
677 Research was partly conducted during a research visit of JJRA to St Andrews University, hosted
678 by OEG and funded by PRX14/00611 mobility grant from the Spanish Ministry of Education,
679 Culture and Sports. OEG was supported by MASTS (the Marine Alliance for Science and
680 Technology for Scotland). We thank three anonymous reviewers for constructive criticism.
681

682 **References**

683

684 Anderson, E.C. & Thompson, E.A. (2002) A model-based method for identifying species hybrids
685 using multilocus genetic data. *Genetics*, **160**, 1217–1229.
686 Bonin, A., Bellemain, E., Eidesen, P.B., Pompanon, F., Brochmann, C. & Taberlet, P. (2004)
687 How to track and assess genotyping errors in population genetics studies. *Molecular*
688 *Ecology*, **13**, 3261–3273.
689 Buchan, J.C., Archie, E.A., Van Horn, R.C., Moss, C.J. & Alberts, S.C. (2005) Locus effects and
690 sources of error in noninvasive genotyping. *Molecular Ecology Notes*, **5**, 680–683.
691 Burczyk, J., Adams, W.T., Birkes, D.S. & Chybicki, I.J. (2006) Using genetic markers to directly
692 estimate gene flow and reproductive success parameters in plants on the basis of naturally
693 regenerated seedlings. *Genetics*, **173**, 363–372.
694 Carlsson, J. (2008) Effects of microsatellite null alleles on assignment testing. *Journal of*
695 *Heredity*, **99**, 616–623.
696 Chapuis, M.P. & Estoup, A. (2007) Microsatellite null alleles and estimation of population
697 differentiation. *Molecular Biology and Evolution*, **24**, 621–631.
698 Chybicki, I.J. & Burczyk, J. (2009) Simultaneous estimation of null alleles and inbreeding
699 coefficients. *Journal of Heredity*, **100**, 106–113.
700 Corander, J., Marttinen, P., Sirén, J. & Tang, J. (2008) Enhanced Bayesian modelling in BAPS
701 software for learning genetic structures of populations. *BMC Bioinformatics*, **9**, 539.
702 Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. & Solignac, M. (1999) New methods employing
703 multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*,
704 **153**, 1989–2000.
705 Dakin, E.E. & Avise, J.C. (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**,
706 504–509.
707 Dawson, K.J. & Belkhir, K. (2001) A Bayesian approach to the identification of panmictic
708 populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.
709 Durand, E., Jay, F., Gaggiotti, O.E. & François, O. (2009) Spatial inference of admixture
710 proportions and secondary contact zones. *Molecular Biology and Evolution*, **26**, 1963–1973.
711 Ellstrand, N.C. (1992) Gene flow by pollen: implications for plant conservation genetics. *Oikos*,
712 **63**, 77–86.

713 Ennos, R.A. (1994) Estimating the relative rates of pollen and seed migration among plant
714 populations. *Heredity*, **72**, 250–259.
715 Falush, D., Stephens, M. & Pritchard, J.K. (2003) Inference of population structure using
716 multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–
717 1587.
718 Faubet, P. & Gaggiotti, O.E. (2008) A new Bayesian method to identify the environmental
719 factors that influence recent migration. *Genetics*, **178**, 1491–1504.
720 Faubet, P., Waples, R. & Gaggiotti, O.E. (2007) Evaluating the performance of a multilocus
721 Bayesian method for the estimation of migration rates. *Molecular Ecology*, **16**, 1149–1166.
722 François, O., Ancelet, S. & Guillot, G. (2006) Bayesian clustering using hidden Markov random
723 fields in spatial population genetics. *Genetics*, **174**, 805–816.
724 Gaggiotti, O.E., Brooks, S.P., Amos, W. & Harwood, J. (2004) Combining demographic,
725 environmental and genetic data to test hypotheses about colonization events in
726 metapopulations. *Molecular Ecology*, **13**, 811–825.
727 Goto, S., Shimatani, K., Yoshimaru, H. & Takahashi, Y. (2006) Fat-tailed gene flow in the
728 dioecious canopy tree species *Fraxinus mandshurica* var. *japonica* revealed by
729 microsatellites. *Molecular Ecology*, **15**, 2985–2996.
730 Guillot, G., Estoup, A., Mortier, F. & Cosson, J.F. (2005) A spatial statistical model for
731 landscape genetics. *Genetics*, **170**, 1261–1280.
732 Hoffman, J.I. & Amos, W. (2005) Microsatellite genotyping errors: detection approaches,
733 common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.
734 Kalinowski, S.T. & Taper, M.L. (2006) Maximum likelihood estimation of the frequency of null
735 alleles at microsatellite loci. *Conservation Genetics*, **7**, 991–995.
736 Kalinowski, S.T., Taper, M.L. & Marshall, T.C. (2007) Revising how the computer program
737 CERVUS accommodates genotyping error increases success in paternity assignment.
738 *Molecular Ecology*, **16**, 1099–1106.
739 Kremer, A., Ronce, O., Robledo-Arnuncio, J.J., Guillaume, F., Bohrer, G., Nathan, R., Bridle,
740 J.R., Gomulkiewicz, R., Klein, E.K., Ritland, K., Kuparinen, A., Gerber, S. & Schueler, S.
741 (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change.
742 *Ecology Letters*, **15**, 378–392.
743 Laikre, L., Schwartz, M.K., Waples, R.S., Ryman, N. & GeM, W.G. (2010) Compromising

genetic diversity in the wild: unmonitored large-scale release of plants and animals. *Trends in Ecology and Evolution*, **25**, 520–529.

Manel, S., Gaggiotti, O.E. & Waples, R.S. (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, **20**, 136–142.

Mills, L.S. & Allendorf, F.W. (1996) The one-migrant-per-generation rule in conservation and management. *Conservation Biology*, **10**, 1509–1518.

Moran, E. V & Clark, J.S. (2011) Estimating seed and pollen movement in a monoecious plant: a hierarchical Bayesian approach integrating genetic and ecological data. *Molecular Ecology*, **20**, 1248–1262.

Morin, P.A., Leduc, R.G., Archer, F.I., Martien, K.K., Huebinger, R., Bickham, J.W. & Taylor, B.L. (2009) Significant deviations from Hardy-Weinberg equilibrium caused by low levels of microsatellite genotyping errors. *Molecular Ecology Resources*, **9**, 498–504.

Paetkau, D., Calvert, W., Stirling, I. & Strobeck, C. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular ecology*, **4**, 347–54.

Pella, J. & Masuda, M. (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fishery Bulletin*, **99**, 151–167.

Pella, J. & Masuda, M. (2006) The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 576–596.

Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews. Genetics*, **6**, 847–859.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rannala, B. & Mountain, J.L. (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 9197–9201.

Robledo-Arnuncio, J.J. (2012) Joint estimation of contemporary seed and pollen dispersal rates among plant populations. *Molecular Ecology Resources*, **12**, 299–311.

Robledo-Arnuncio, J.J., Klein, E.K., Muller-Landau, H.C. & Santamaría, L. (2014) Space, time and complexity in plant dispersal ecology. *Movement Ecology*, **2**, 16.

Selkoe, K.A. & Toonen, R.J. (2006) Microsatellites for ecologists: a practical guide to using and

775 evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.
776 Sieberts, S.K., Wijsman, E.M. & Thompson, E.A. (2002) Relationship inference from trios of
777 individuals, in the presence of typing error. *American Journal of Human Genetics*, **70**, 170–
778 180.
779 Slavov, G.T., Howe, G.T., Gyaourova, a. V., Birkes, D.S. & Adams, W.T. (2005) Estimating
780 pollen flow using SSR markers and paternity exclusion: accounting for mistyping.
781 *Molecular Ecology*, **14**, 3109–3121.
782 Unger, G.M., Vendramin, G.G. & Robledo-Arnuncio, J.J. (2014) Estimating exotic gene flow
783 into native pine stands: zygotic vs. gametic components. *Molecular Ecology*, **23**, 5435–
784 5447.
785 Wang, J. (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**,
786 1963–1979.
787 Wang, J. (2014) Estimation of migration rates from marker-based parentage analysis. *Molecular*
788 *Ecology*, **23**, 3191–3213.
789 Waples, R.S. & Gaggiotti, O. (2006) What is a population? An empirical evaluation of some
790 genetic methods for identifying the number of gene pools and their degree of connectivity.
791 *Molecular Ecology*, **15**, 1419–1439.
792 Wilson, G.A. & Rannala, B. (2003) Bayesian inference of recent migration rates using multilocus
793 genotypes. *Genetics*, **163**, 1177–1191.
794

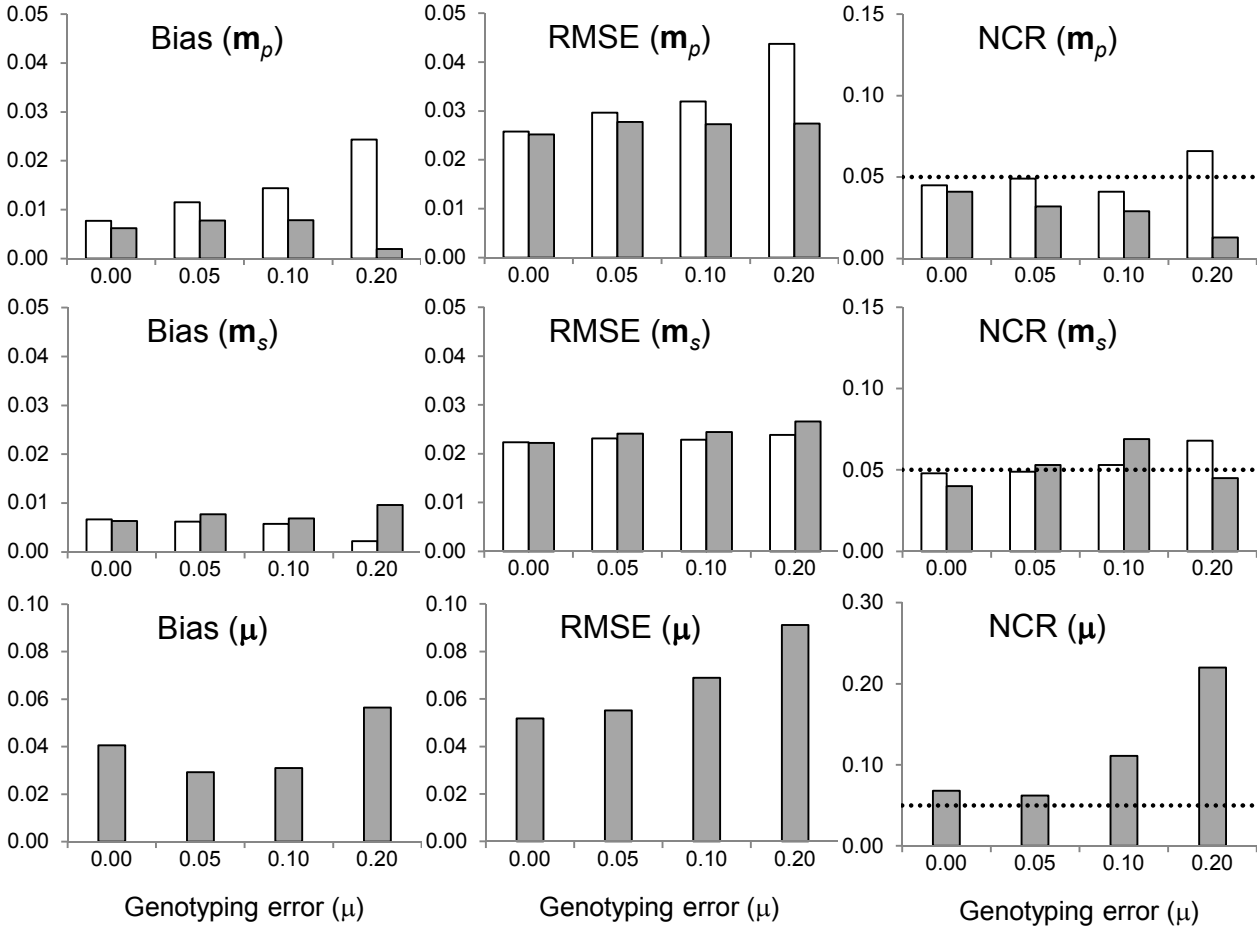


**Fig. 1**. Effect of genotyping error rate (μ) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained when migration is non-null ($m_{pi} = m_{si} = 0.05$) by either ignoring (white bars) or jointly estimating (grey bars) μ. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $F_{ST} = 0.1$, $L = 10$ loci (all with equal μ), $k_l = 10$ alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

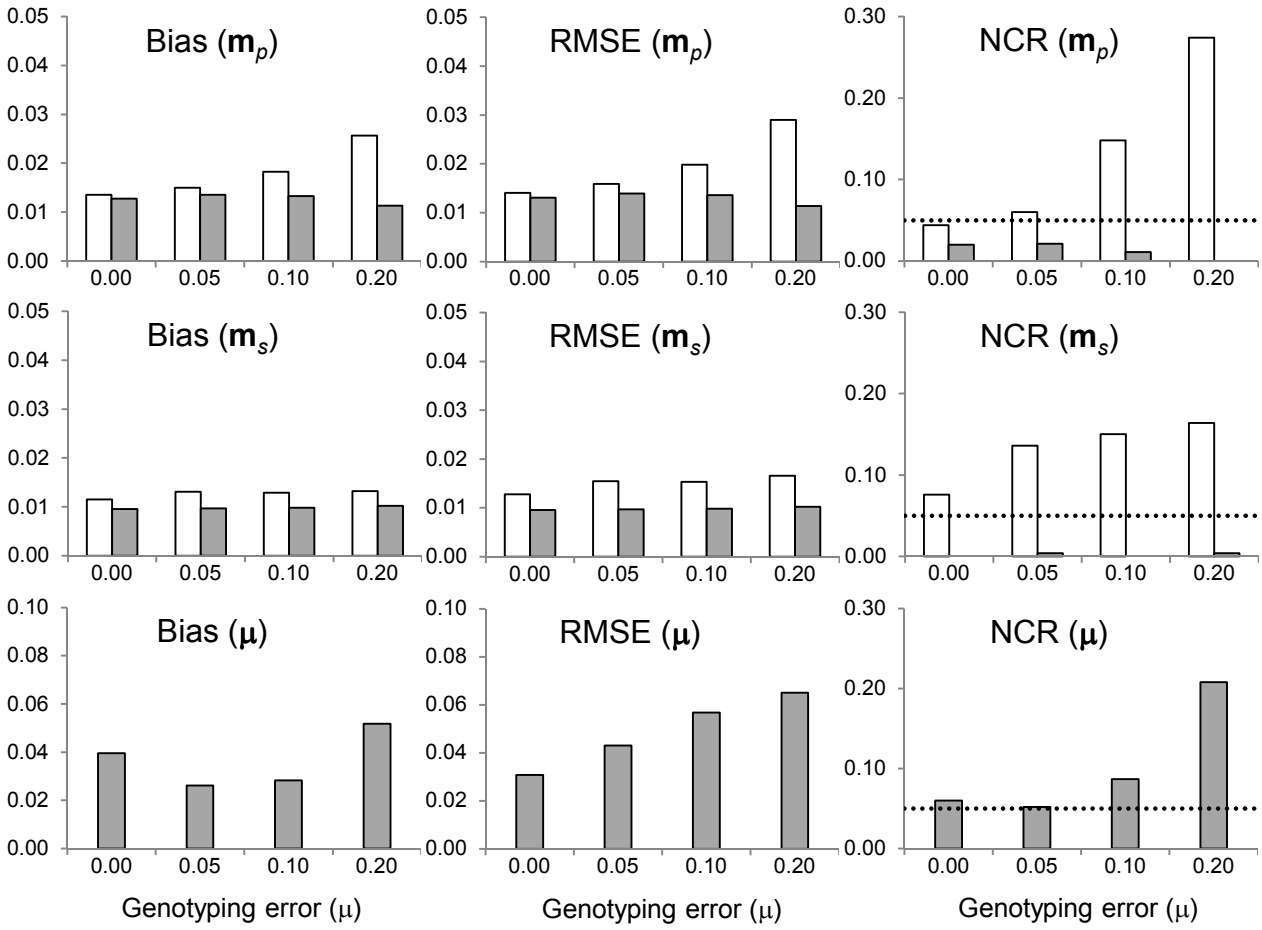


**Fig. 2.** Effect of genotyping error rate (μ) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained in absence of migration ($m_{pi} = m_{si} = 0$) by either ignoring (white bars) or jointly estimating (grey bars) μ. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $F_{ST} = 0.1$, $L = 10$ loci (all with equal μ), $k_l = 10$ alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

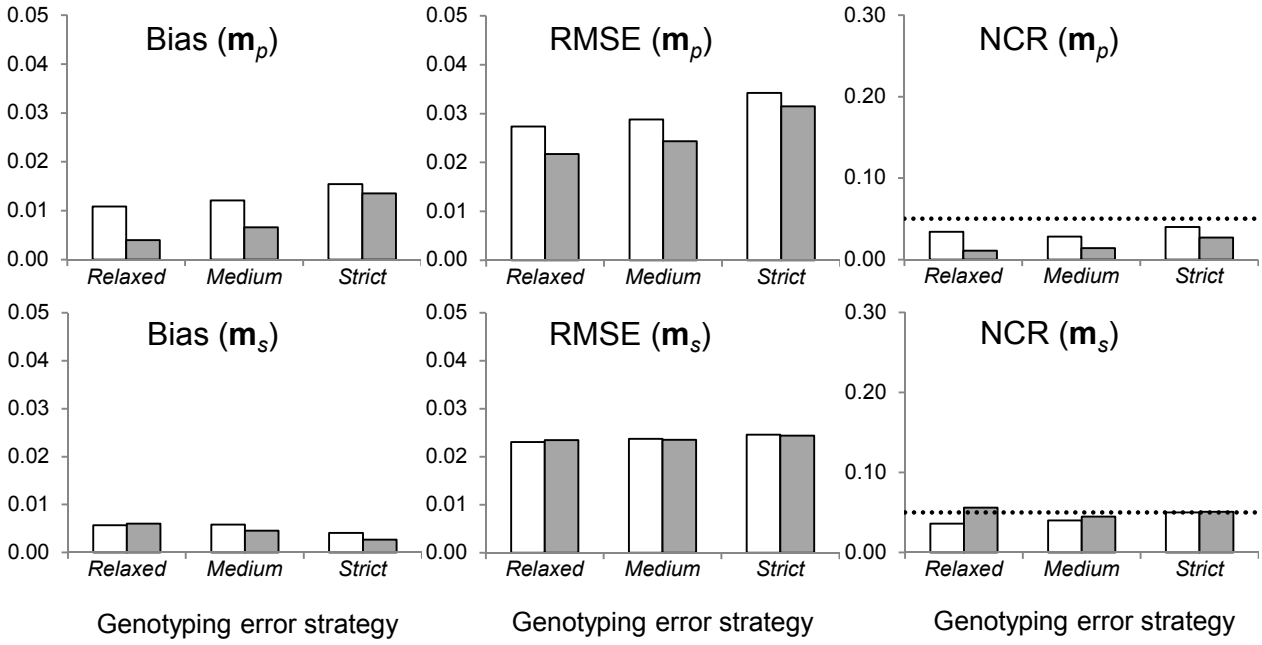


**Fig. 3.** Effect of including or excluding marker loci with variable genotyping error rate (μ) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained either ignoring (white bars) or jointly estimating (grey bars) μ. Ten loci were assumed: five with μ = 0, three with μ = 0.05 and two with μ = 0.1. The *relaxed* strategy used all loci, while *Medium* and *Strict* discarded those with μ ≥ 0.1 and μ ≥ 0.05, respectively. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $m_{pi} = m_{si} = 0.05$, $F_{ST} = 0.1$, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

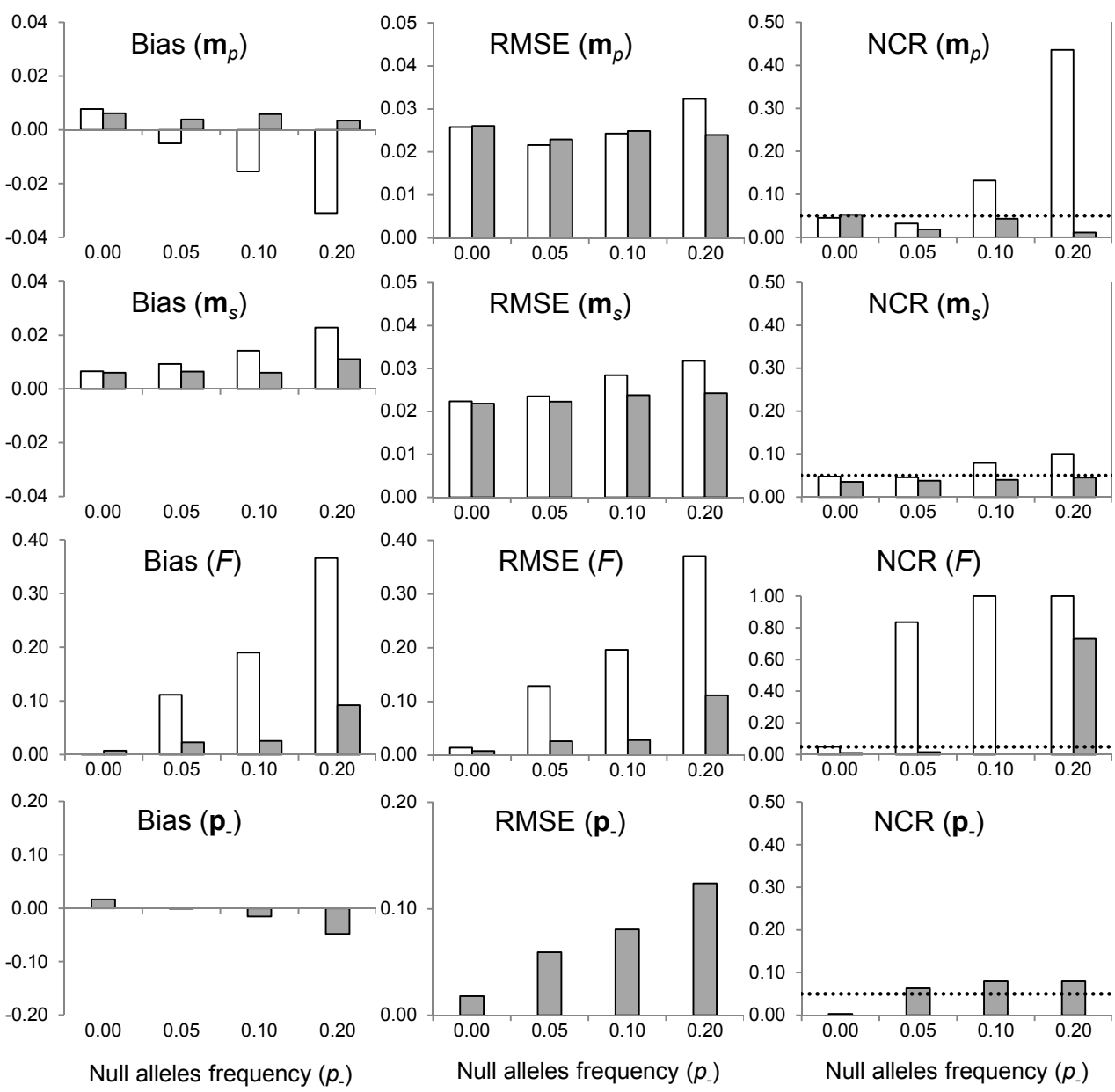


**Fig. 4.** Effect of null alleles frequency ($p_-$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and population inbreeding ($F$) estimates obtained by either ignoring (white bars) or jointly estimating (grey bars) $p_-$. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $m_{pi} = m_{si} = 0.05$, $F_{ST} = 0.1$, $F = 0$, $L = 5$ (for $p_- = 0.05$ scenario) or 10 (for $p_- \neq 0.05$ scenarios) loci, $k_l = 19$ (for $p_- = 0.05$), 10 (for $p_- = 0$), 9 (for $p_- = 0.1$), or 8 (for $p_- = 0.2$) non-null alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

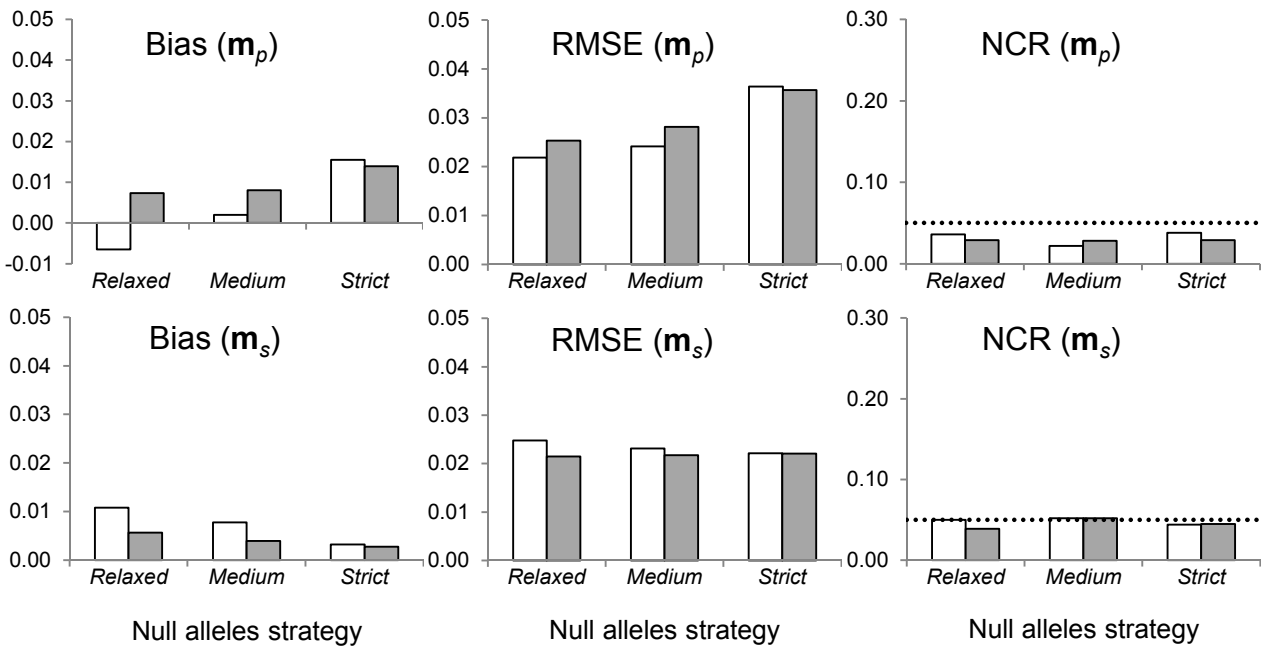


**Fig. 5**. Effect of including or excluding marker loci with variable null allele frequencies ($p_-$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained either ignoring (white bars) or jointly estimating (grey bars) $p_-$. Ten loci were assumed: five with $p_- = 0$, three with $p_- = 0.1$ and two with $p_- = 0.2$ (with $k_l = 10$, 9, and 8 non-null alleles, respectively). The *relaxed* strategy used all loci, while *Medium* and *Strict* discarded those with $p_- \geq 0.2$ and $p_- \geq 0.1$, respectively. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $m_{pi} = m_{si} = 0.05$, $F_{ST} = 0.1$, $F = 0$, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

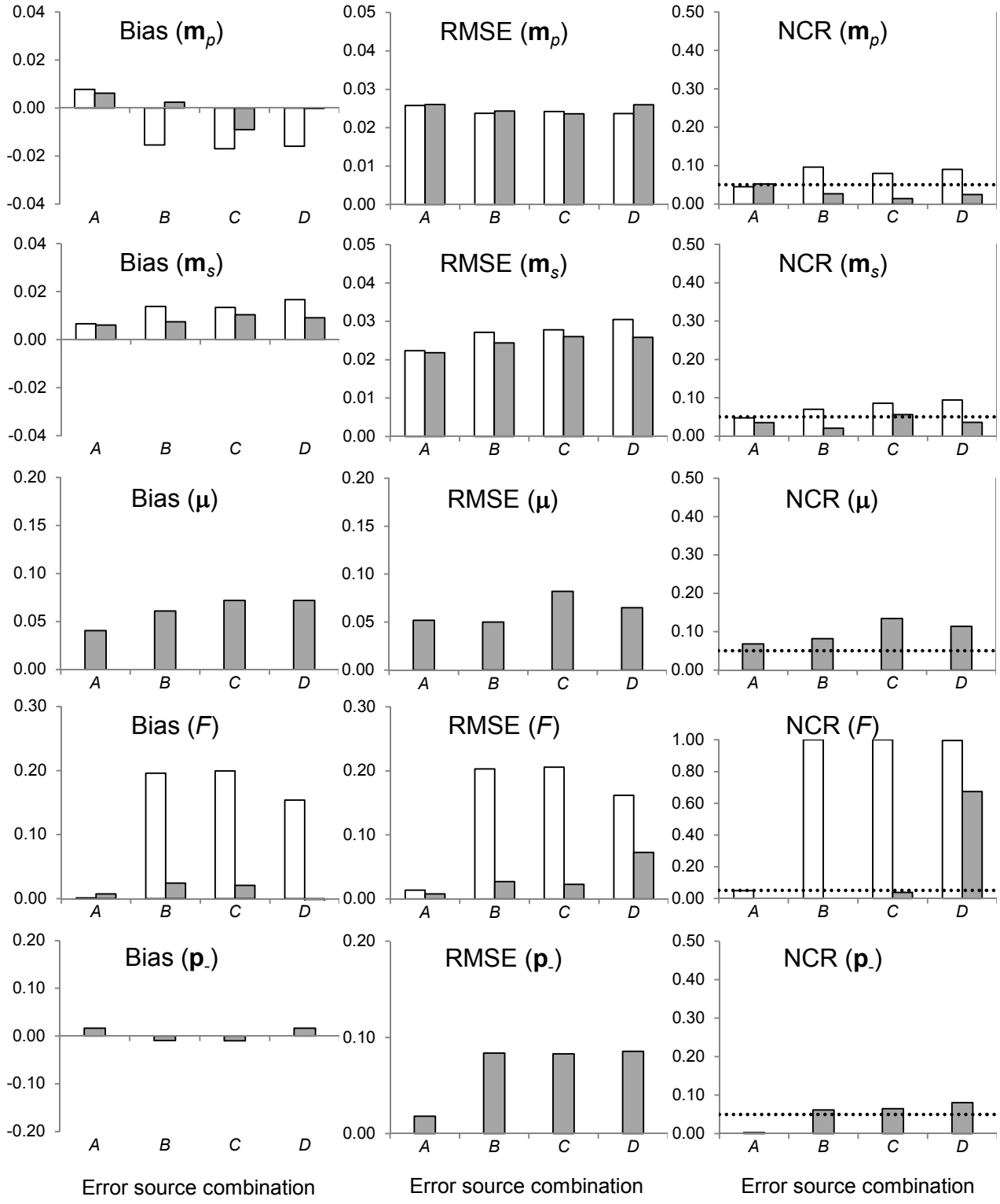


**Fig. 6**. Combined effect of genotyping error rate (μ) and null alleles frequency ($p_-$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and population inbreeding ($F$) estimates obtained by either

ignoring (white bars) or jointly estimating (grey bars) μ and $p_{–}$. Assumed scenarios were: *A* ($F$ = μ = $p_{–}$ = 0), *B* ($F$ = 0, μ = 0.05, $p_{–}$ = 0.1), *C* ($F$ = 0, μ = 0.1, $p_{–}$ = 0.1), and *D* ($F$ = μ = $p_{–}$ = 0.1). RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I$ = 2 external populations, $m_{pi}$ = $m_{si}$ = 0.05, $F_{ST}$ = 0.1, $L$ = 10 loci (all with equal μ and $p_{–}$), $k_l$ = 10 (in *A*) or 9 (in *B*, *C* and *D*) non-null alleles/locus, sample sizes of $D$ = 100 offspring and $Q_i$ = 100 adults/population.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

**Appendix S1** *Offspring genotypic likelihoods with genotyping errors and null alleles*

**Appendix S2** *Adult genotypic likelihoods with genotyping errors and null alleles*

**Appendix S3** *Details of the MCMC algorithm*

**Appendix S4** *Monte Carlo analysis of method performance*

**Figure S1** *Effect of genotyping errors on migration estimates with 20 loci and 5alleles/locus*

**Figure S2** *Effect of genotyping errors on migration estimates with 5 loci and 20 alleles/locus*

**Figure S3** *Effect of null alleles on migration estimates in absence of seed migration*

**Figure S4** *Effect of null alleles on migration estimates in absence of pollen migration*

**Figure S5** *Effect of null alleles and population inbreeding on migration estimates*

**Figure S6** *Effect of population inbreeding (without null alleles) on migration estimates*

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

**Appendix S1** *Offspring genotypic likelihoods with genotyping errors and null alleles*

Offspring genotypic likelihoods with population inbreeding and null alleles, but without genotyping errors, are presented in equation 3 of the main text. Here, we derive the genotypic likelihood of an offspring $d$ at locus $l$, $\Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \beta_l, \mu_l)$, in presence of all population inbreeding, null alleles and genotyping errors. We recall that our model assumes that any true non-null allele (including the two homologous alleles in a genotype, irrespective of whether they are equal or not) is independently mistaken for any of the other $k_l - 1$ alleles with equal probability $\mu_l/(k_l - 1)$, for adult and offspring genotypes alike. For the sake of illustration, we first assume there are typing errors and population inbreeding, but not null alleles, in which case we can drop $\beta_l$ (because without null alleles it is not necessary to estimate the proportion of non-null missing data, which is directly observed) and have

$$\Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \mu_l) = \begin{cases} (1-F_i)A + F_i\left[p_{ila_1}(1-\mu_l)^2 + \dfrac{\mu_l^2(1-p_{ila_1})}{(k_l-1)^2}\right] & \text{if } i = j \text{ and } a_1 = a_2 \\ (1-F_i)B + F_i\left[\dfrac{2\mu_l(1-\mu_l)(p_{ila_1}+p_{ila_2})}{k_l-1} + \dfrac{2\mu_l^2(1-p_{ila_1}-p_{ila_2})}{(k_l-1)^2}\right] & \text{if } i = j \text{ and } a_1 \neq a_2 \\ C & \text{if } i \neq j \text{ and } a_1 = a_2 \\ D & \text{if } i \neq j \text{ and } a_1 \neq a_2 \end{cases} \quad \text{eqn S1}$$

where $A$, $B$, $C$ and $D$ are as follows:

$$A = (1-\mu_l)^2 p_{ila_1}^2 + \frac{2\mu_l(1-\mu_l)}{k_l-1} p_{ila_1} \sum_{b \neq a_1} p_{ilb} + \frac{\mu_l^2}{(k_l-1)^2} \sum_{b,c \neq a_1} \delta_{bc} p_{ilb} p_{ilc} =$$

$$= (1-\mu_l)^2 p_{ila_1}^2 + \frac{2\mu_l(1-\mu_l)}{k_l-1} p_{ila_1}(1-p_{ila_1}) + \frac{\mu_l^2}{(k_l-1)^2} \sum_{b,c \neq a_1} \delta_{bc} p_{ilb} p_{ilc}$$

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

$$B = 2(1-\mu_l)^2 p_{ila_1} p_{ila_2} + \frac{2\mu_l(1-\mu_l)}{k_l - 1}\left( p_{ila_1} \sum_{b \neq a_2} p_{ilb} + p_{ila_2} \sum_{b \neq a_1} p_{ilb} \right) +$$

$$+ \frac{\mu_l^2}{(k_l-1)^2}\left( p_{ila_1} \sum_{b \neq a_1} \delta_{a_2 b} p_{ilb} + p_{ila_2} \sum_{b \neq a_2} \delta_{a_1 b} p_{ilb} + 2 \sum_{b,c \neq a_1, a_2} \delta_{bc} p_{ilb} p_{ilc} \right) =$$

$$= 2(1-\mu_l)^2 p_{ila_1} p_{ila_2} + \frac{2\mu_l(1-\mu_l)}{k_l - 1}\left[ p_{ila_1}(1 - p_{ila_2}) + p_{ila_2}(1 - p_{ila_1}) \right] +$$

$$+ \frac{\mu_l^2}{(k_l-1)^2}\left[ 2(p_{ila_1} + p_{ila_2})(1 - p_{ila_1} - p_{ila_2}) + 2 p_{ila_1} p_{ila_2} + 2 \sum_{b,c \neq a_1, a_2} \delta_{bc} p_{ilb} p_{ilc} \right]$$

$$\text{with } \delta_{xy} = \begin{cases} 1 & if \quad x = y \\ 2 & if \quad x \neq y \end{cases}$$

$$C = (1-\mu_l)^2 p_{ila_1} p_{jla_1} + \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[ p_{ila_1} \sum_{b \neq a_1} p_{jlb} + p_{jla_1} \sum_{b \neq a_1} p_{ilb} \right] + \frac{\mu_l^2}{(k_l-1)^2} \sum_{b,c \neq a_1} p_{ilb} p_{jlc} =$$

$$= (1-\mu_l)^2 p_{ila_1} p_{jla_1} + \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[ p_{ila_1}(1 - p_{jla_1}) + p_{jla_1}(1 - p_{ila_1}) \right] + \frac{\mu_l^2}{(k_l-1)^2}(1 - p_{ila_1})(1 - p_{jla_1})$$

$$D = (1-\mu_l)^2 (p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1}) + \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[ p_{ila_1} \sum_{b \neq a_2} p_{jlb} + p_{ila_2} \sum_{b \neq a_1} p_{jlb} + p_{jla_1} \sum_{b \neq a_2} p_{ilb} + p_{jla_2} \sum_{b \neq a_1} p_{ilb} \right] +$$

$$+ \frac{\mu_l^2}{(k_l-1)^2}\left( p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1} + p_{ila_1} \sum_{b \neq a_1, a_2} p_{jlb} + p_{ila_2} \sum_{b \neq a_1, a_2} p_{jlb} + p_{jla_1} \sum_{b \neq a_1, a_2} p_{ilb} + p_{jla_2} \sum_{b \neq a_1, a_2} p_{ilb} + 2 \sum_{b,c \neq a_1, a_2} p_{ilb} p_{jlc} \right) =$$

$$= (1-\mu_l)^2 (p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1}) + \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[ p_{ila_1}(1 - p_{jla_2}) + p_{ila_2}(1 - p_{jla_1}) + p_{jla_1}(1 - p_{ila_2}) + p_{jla_2}(1 - p_{ila_1}) \right] +$$

$$+ \frac{\mu_l^2}{(k_l-1)^2}\left[ p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1} + (p_{ila_1} + p_{ila_2})(1 - p_{jla_1} - p_{jla_2}) + (p_{jla_1} + p_{jla_2})(1 - p_{ila_1} - p_{ila_2}) + 2 \sum_{b,c \neq a_1, a_2} p_{ilb} p_{jlc} \right]$$

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

We now consider both null alleles and genotyping errors, assuming that null alleles are not mistaken by non-null alleles and *vice versa*. We then have

$$\Pr_{ij}(X_{dl} \mid \mathbf{p}, F_i, \beta_l, \mu_l) = \begin{cases} (1-\beta_l)\left( (1-F_i)A' + F_i\left[ p_{ila_1}(1-\mu_l)^2 + \dfrac{\mu_l^2(1-p_{ila_1}-p_{il-})}{(k_l-1)^2} \right] \right) & \text{if } i=j \text{ and } a_1=a_2 \\ (1-\beta_l)\left( (1-F_i)B' + F_i\left[ \dfrac{2\mu_l(1-\mu_l)(p_{ila_1}+p_{ila_2})}{k_l-1} + \dfrac{2\mu_l^2(1-p_{ila_1}-p_{ila_2}-p_{il-})}{(k_l-1)^2} \right] \right) & \text{if } i=j \text{ and } a_1\neq a_2 \\ (1-\beta_l)C' & \text{if } i\neq j \text{ and } a_1=a_2 \\ (1-\beta_l)D' & \text{if } i\neq j \text{ and } a_1\neq a_2 \\ \beta_l + (1-\beta_l)\left[(1-F_i)p_{il-}^2 + F_i p_{il-}\right] & \text{if } i=j \text{ and missing data} \\ \beta_l + (1-\beta_l)p_{il-}p_{jl-} & \text{if } i\neq j \text{ and missing data} \end{cases}$$

eqn S2

where *A'*, *B'*, *C'* and *D'* are as follows:

$$A' = (1-\mu_l)^2 p_{ila_1}^2 + 2(1-\mu_l)p_{ila_1}p_{il-} + \frac{2\mu_l(1-\mu_l)}{k_l-1}p_{ila_1}(1-p_{ila_1}-p_{il-}) + \frac{2\mu}{k_l-1}p_{il-}(1-p_{ila_1}-p_{il-}) +$$

$$+ \frac{\mu_l^2}{(k_l-1)^2}\sum_{\substack{b,c\neq a_1 \\ b,c\neq null}} \delta_{bc}p_{ilb}p_{ilc} =$$

$$= (1-\mu_l)^2 p_{ila_1}^2 + 2(1-\mu_l)p_{ila_1}p_{il-} + \frac{2\mu}{k_l-1}(1-p_{ila_1}-p_{il-})\left[p_{ila_1}(1-\mu_l)+p_{il-}\right] + \frac{\mu_l^2}{(k_l-1)^2}\sum_{\substack{b,c\neq a_1 \\ b,c\neq null}} \delta_{bc}p_{ilb}p_{ilc}$$

$$B' = 2(1-\mu_l)^2 p_{ila_1}p_{ila_2} + \frac{2\mu_l(1-\mu_l)}{k_l-1}\left[p_{ila_1}(1-p_{ila_2}-p_{il-}) + p_{ila_2}(1-p_{ila_1}-p_{il-})\right] +$$

$$+ \frac{2\mu_l^2}{(k_l-1)^2}\left[ p_{ila_1}p_{ila_2} + (p_{ila_1}+p_{ila_2})(1-p_{ila_1}-p_{ila_2}-p_{il-}) + \sum_{\substack{b,c\neq a_1,a_2 \\ b,c\neq null}} \delta_{bc}p_{ilb}p_{ilc} \right]$$

$$\text{with } \delta_{xy} = \begin{cases} 1 & \text{if } x=y \\ 2 & \text{if } x\neq y \end{cases}$$

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

$$C' = (1-\mu_l)^2 p_{ila_1} p_{jla_1} + (1-\mu_l)(p_{il-} p_{jla_1} + p_{ila_1} p_{jl-}) +$$
$$+ \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[p_{ila_1}(1 - p_{jla_1} - p_{jl-}) + p_{jla_1}(1 - p_{ila_1} - p_{il-}) + p_{il-}(1 - p_{jla_1} - p_{jl-}) + p_{jl-}(1 - p_{ila_1} - p_{il-})\right] +$$
$$+ \frac{\mu_l^2}{(k_l - 1)^2} \sum_{b,c \neq a_1, null} p_{ilb} p_{jlc} =$$
$$= (1-\mu_l)^2 p_{ila_1} p_{jla_1} + (1-\mu_l)(p_{il-} p_{jla_1} + p_{ila_1} p_{jl-}) +$$
$$+ \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[(p_{ila_1} + p_{il-})(1 - p_{jla_1} - p_{jl-}) + (p_{jla_1} + p_{jl-})(1 - p_{ila_1} - p_{il-})\right] +$$
$$+ \frac{\mu_l^2}{(k_l - 1)^2}\left[(1 - p_{ila_1} - p_{il-})(1 - p_{jla_1} - p_{jl-})\right]$$

$$D' = (1-\mu_l)^2 (p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1}) +$$
$$+ \frac{\mu_l(1-\mu_l)}{k_l - 1}\begin{bmatrix}(p_{ila_1} + p_{ila_2})(1 - p_{jla_1} - p_{jla_2} - p_{jl-}) + (p_{jla_1} + p_{jla_2})(1 - p_{ila1} - p_{ila_2} - p_{il-}) + \\ + 2p_{ila1} p_{jla_1} + 2p_{ila_2} p_{jla_2}\end{bmatrix} +$$
$$+ \frac{\mu_l^2}{(k_l - 1)^2}\begin{bmatrix}p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1} + (p_{ila_1} + p_{ila_2})(1 - p_{jla_1} - p_{jla_2} - p_{jl-}) + \\ + (p_{jla_1} + p_{jla_2})(1 - p_{ila1} - p_{ila_2} - p_{il-}) + \sum_{c,d \neq a_1, a_2, null} 2p_{ilc} p_{jld}\end{bmatrix} =$$
$$= (1-\mu_l)^2 (p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1}) +$$
$$+ \frac{\mu_l(1-\mu_l)}{k_l - 1}\left[p_{ila_1}(1 - p_{jla_2}) + p_{ila_2}(1 - p_{jla_1}) + p_{jla_1}(1 - p_{ila_2}) + p_{jla_2}(1 - p_{ila_1})\right] +$$
$$+ \frac{\mu_l^2}{(k_l - 1)^2}\left[p_{ila_1} p_{jla_2} + p_{ila_2} p_{jla_1} + (1 - p_{il-})(1 - p_{jla_1} - p_{jla_2} - p_{jl-}) + (1 - p_{jl-})(1 - p_{ila_1} - p_{ila_2} - p_{il-})\right]$$

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

**Appendix S2** *Adult genotypic likelihoods with genotyping errors and null alleles*

The joint likelihood of the sample of $Q_i$ adult genotypes observed at locus $l$ in population $i$ is

$$\Pr(\mathbf{Y}_{il} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l) = \prod_{q=1}^{Q_i} \Pr(Y_{iql} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l),$$

where $Y_{iql}$ is the diploid genotype of adult $q$ from population $i$ at locus $l$. The likelihood of $Y_{iql}$ depends on whether any allele is observed or not (i.e. missing data) and on whether the two eventually observed homologous alleles (denoted $a_1$ and $a_2$) are equal ($Y_{iql}$ homozygous) or not ($Y_{iql}$ heterozygous):

$$\Pr(Y_{iql} \mid \mathbf{p}_{il}, F_i^*, \beta_l^*, \mu_l) = \begin{cases} (1-F_i^*)A' + F_i^* \left[ p_{ila_1}(1-\mu_l)^2 + \dfrac{\mu_l^2(1-p_{ila_1}-p_{il-})}{(k_l-1)^2} \right] & \text{if} \quad a_1 = a_2 \\ (1-F_i^*)B' + F_i^* \left[ \dfrac{2\mu_l(1-\mu_l)(p_{ila_1}+p_{ila_2})}{k_l-1} + \dfrac{2\mu_l^2(1-p_{ila_1}-p_{ila_2}-p_{il-})}{(k_l-1)^2} \right] & \text{if} \quad a_1 \neq a_2 \\ \beta_l^* + (1-\beta_l^*)\left[(1-F_i^*)p_{il-}^2 + F_i^* p_{il-}\right] & \text{if} \quad \textit{missing data} \end{cases}$$

eqn S3

where *A'* and *B'* are defined as in Appendix S1.

**Appendix S3** *Details of the MCMC algorithm*

We used the Metropolis-Hastings (MH) algorithm (Metropolis *et al.* 1953; Hastings 1970) to approximate numerically the posterior probability density of model parameters. Our implementation of the algorithm involved eight steps at each MH iteration, with a different parameter vector being potentially modified at each step, as described below. Initial values for **p** were set at their observed values in parental samples (with $p_{il-} = 0$). Initial values for **m** and **F** were set at their maximum-likelihood estimates (using equation 1), given fixed **p** at observed allelic frequencies in parental samples (with $p_{il-} = 0$) and fixed **μ** at zero. Initial values for **μ** and **F*** were set at zero, while those of **β** and **β*** were set at the observed rates of missing data in offspring and adults samples, respectively.

*Updating offspring proportions (***m***)*

We updated local dispersal and/or seed and pollen immigration rates in a single step. The vector **m** of size $2I+1$ was obtained by concatenating the current values of vectors $\mathbf{m}_p$ and $\mathbf{m}_s$ and the current local dispersal rate $m_0$. Note that the elements of **m** add up to one, as defined in our model. We randomly choose two different elements $i$ and $j$ of vector **m**, and propose a new value $m_i' = u$, where $u \sim U\,[\max(0, m_i - e_m), \min(m_i + e_m, m_i + m_j)]$ and $e_m$ is some incremental value that is tuned to obtain reasonable acceptance rates. We then set $m_j' = m_i + m_j - m_i'$. The updated vector **m'** (incorporating the updated values $m_i'$ and $m_j'$) is accepted with probability

$$\alpha(\mathbf{m}'|\mathbf{m}) = \min\left(1, \frac{\Pr(\mathbf{X}|\mathbf{m}',\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\,f(\mathbf{m}')\,q(\mathbf{m},\mathbf{m}')}{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\,f(\mathbf{m})\,q(\mathbf{m}',\mathbf{m})}\right),$$

where $q(\mathbf{m}, \mathbf{m}')$ is the uniform density described above and $q(\mathbf{m}', \mathbf{m})$ is the uniform density corresponding to the reverse move, namely $U[\max(0,\ m_i' - e_m), \min(m_i' + e_m,\ m_i' + m_j')]$.

*Updating offspring population inbreeding coefficients (***F***)*

We updated the offspring inbreeding coefficients of all populations at every MH iteration, one at a time. For each population *i*, we propose a new value $F_i^{'} = u$, where $u \sim U$ [max($F_{min}$, $F_i - e_F$), min(1, $F_i + e_F$)], $e_F$ is some incremental value, and $F_{min} = \max(-1, -p_{min}/(1-p_{min}))$, with $p_{min}$ being the minimum of allele frequencies across loci in population *i* (Faubet & Gaggiotti 2008). Let **F'** be the vector of inbreeding coefficients **F** with element $F_i$ replaced with $F_i^{'}$. This move is accepted with probability

$$\alpha(\mathbf{F'}\,|\,\mathbf{F}) = \min\left(1, \frac{\Pr(\mathbf{X}\,|\,\mathbf{m},\mathbf{p},\mathbf{F'},\boldsymbol{\beta},\boldsymbol{\mu})\, f(\mathbf{F'})\, q(\mathbf{F},\mathbf{F'})}{\Pr(\mathbf{X}\,|\,\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\, f(\mathbf{F})\, q(\mathbf{F'},\mathbf{F})}\right),$$

where $q$(**F**, **F'**) is the uniform density described above and $q$(**F'**, **F**) is the reverse move uniform density $U$[max($F_{min}$, $F_i^{'} - e_F$), min(1, $F_i^{'} + e_F$)].

*Updating population frequencies of non-null alleles (***p***)*

We updated the population allelic frequencies of all populations and loci at every MH iteration, one population and locus at a time. Null alleles were updated separately (see next section), so that acceptance rates could be tuned independently. For each population *i* and locus *l*, we randomly choose two of the $k_l$ non-null alleles, *a* and *b*, and propose $p_{ila}^{'} = u$, where $u \sim U$ [max(0, $p_{ila} - e_p$), min($p_{ila} + e_p$, $p_{ila} + p_{ilb}$)] and $e_p$ is some incremental value. We then set $p_{ilb}^{'} = p_{ila} + p_{ilb} - p_{ila}^{'}$ and accept the move with probability

$$\alpha(\mathbf{p'}\,|\,\mathbf{p}) = \min\left(1, \frac{\Pr(\mathbf{X}\,|\,\mathbf{m},\mathbf{p'},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\Pr(\mathbf{Y}\,|\,\mathbf{p'},\mathbf{F}^{*},\boldsymbol{\beta}^{*},\boldsymbol{\mu})\, f(\mathbf{p'})\, q(\mathbf{p},\mathbf{p'})}{\Pr(\mathbf{X}\,|\,\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\Pr(\mathbf{Y}\,|\,\mathbf{p},\mathbf{F}^{*},\boldsymbol{\beta}^{*},\boldsymbol{\mu})\, f(\mathbf{p})\, q(\mathbf{p'},\mathbf{p})}\right),$$

where **p'** is the vector **p** with elements $p_{ila}$ and $p_{ilb}$ replaced with $p'_{ila}$ and $p'_{ilb}$, $q(\mathbf{p}, \mathbf{p'})$ is the uniform density described above, and $q(\mathbf{p'}, \mathbf{p})$ is the reverse move uniform density $U$ [max(0, $p'_{ila} - e_p$), min($p'_{ila} + e_p$, $p'_{ila} + p'_{ilb}$)].

*Updating population frequencies of null alleles (*$\mathbf{p}_{il-}$*)*

We updated the null allele frequencies at all loci and populations at every MH iteration, one population and locus at a time. For each population $i$ and locus $l$, we propose $p'_{il-} = u$, where $u \sim U$ [max(0, $p_{il-} - e_{p-}$), min(1, $p_{il-} + e_{p-}$)] and $e_{p-}$ is some incremental value. We then set the frequency of each non-null allele $a$ at $p'_{ila} = p_{ila}(1 - p'_{il-})/(1 - p_{il-})$, and accept the move with probability

$$\alpha(\mathbf{p'}|\mathbf{p}) = \min\left(1, \frac{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p'},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\Pr(\mathbf{Y}|\mathbf{p'},\mathbf{F}^*,\boldsymbol{\beta}^*,\boldsymbol{\mu}) f(\mathbf{p'}) q(\mathbf{p},\mathbf{p'})}{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu})\Pr(\mathbf{Y}|\mathbf{p},\mathbf{F}^*,\boldsymbol{\beta}^*,\boldsymbol{\mu}) f(\mathbf{p}) q(\mathbf{p'},\mathbf{p})}\right),$$

where **p'** is the vector **p** with elements $p_{il-}$ and $p_{ila}$ ($a = 1, \ldots, k_l$) replaced with $p'_{il-}$ and $p'_{ila}$, $q(\mathbf{p}, \mathbf{p'})$ is the uniform density described above, and $q(\mathbf{p'}, \mathbf{p})$ is the reverse move uniform density $U$ [max(0, $p'_{il-} - e_{p-}$), min(1, $p'_{il-} + e_{p-}$)].

*Updating offspring non-null missing data rates (*$\boldsymbol{\beta}$*)*

We updated the rate of offspring missing data (not due to null alleles) at all loci at every MH iteration, one locus at a time. For each locus $l$, we propose $\beta'_l = u$, where $u \sim U$ [max(0, $\beta_l - e_\beta$), min(1, $\beta_l + e_\beta$)] and $e_\beta$ is some incremental value. Denote **β'** the vector **β** with element $\beta_l$ replaced with $\beta'_l$, the move is then accepted with probability

$$\alpha(\boldsymbol{\beta}'|\boldsymbol{\beta}) = \min\left(1, \frac{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta}',\boldsymbol{\mu}) f(\boldsymbol{\beta}') q(\boldsymbol{\beta},\boldsymbol{\beta}')}{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}) f(\boldsymbol{\beta}) q(\boldsymbol{\beta}',\boldsymbol{\beta})}\right),$$

where $q(\boldsymbol{\beta}, \boldsymbol{\beta}')$ is the uniform density described above and $q(\boldsymbol{\beta}', \boldsymbol{\beta})$ is the reverse move uniform density $U[\max(0,\ \beta_l' - e_\beta),\ \min(1,\ \beta_l' + e_\beta)]$.

*Updating genotyping error rates (**μ**)*

We updated genotyping error rates of all loci at every MH iteration, one locus at a time. For each locus $l$, we propose $\mu_l' = u$, where $u \sim U\ [\max(0,\ \mu_l - e_\mu),\ \min(1,\ \mu_l + e_\mu)]$ and $e_\mu$ is some incremental value. Denote **μ'** the vector **μ** with element $\mu_l$ replaced with $\mu_l'$, the move is then accepted with probability

$$\alpha(\boldsymbol{\mu}'|\boldsymbol{\mu}) = \min\left(1, \frac{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}') \Pr(\mathbf{Y}|\mathbf{p},\mathbf{F}^*,\boldsymbol{\beta}^*,\boldsymbol{\mu}') f(\boldsymbol{\mu}') q(\boldsymbol{\mu},\boldsymbol{\mu}')}{\Pr(\mathbf{X}|\mathbf{m},\mathbf{p},\mathbf{F},\boldsymbol{\beta},\boldsymbol{\mu}) \Pr(\mathbf{Y}|\mathbf{p},\mathbf{F}^*,\boldsymbol{\beta}^*,\boldsymbol{\mu}) f(\boldsymbol{\mu}) q(\boldsymbol{\mu}',\boldsymbol{\mu})}\right),$$

where $q(\boldsymbol{\mu}, \boldsymbol{\mu}')$ is the uniform density described above and $q(\boldsymbol{\mu}', \boldsymbol{\mu})$ is the reverse move uniform density $U[\max(0,\ \mu_l' - e_\mu),\ \min(1,\ \mu_l' + e_\mu)]$.

*Updating adult population inbreeding coefficients (**F***)*

We updated adult inbreeding coefficients for all populations at every MH iteration, one at a time. For each population $i$, we propose a new value $F_i^{*'} = u$, where $u \sim U\ [\max(F_{min},\ F_i^* - e_{F*}),\ \min(1,\ F_i^* + e_{F*})]$, $e_{F*}$ is some incremental value, and $F_{min}$ defined as when updating **F**. Let **F*'** be the vector of inbreeding coefficients **F*** with element $F_i^*$ replaced with $F_i^{*'}$. This move is accepted with probability

$$\alpha(\mathbf{F}^{*\prime}|\,\mathbf{F}^{*}) = \min\left(1, \frac{\Pr(\mathbf{Y}\,|\,\mathbf{p}, \mathbf{F}^{*\prime}, \boldsymbol{\beta}^{*}, \boldsymbol{\mu}) f(\mathbf{F}^{*\prime}) q(\mathbf{F}^{*}, \mathbf{F}^{*\prime})}{\Pr(\mathbf{Y}\,|\,\mathbf{p}, \mathbf{F}^{*}, \boldsymbol{\beta}^{*}, \boldsymbol{\mu}) f(\mathbf{F}^{*}) q(\mathbf{F}^{*\prime}, \mathbf{F}^{*})}\right),$$

where $q(\mathbf{F}^{*}, \mathbf{F}^{*\prime})$ is the uniform density described above and $q(\mathbf{F}^{*\prime}, \mathbf{F}^{*})$ is the reverse move uniform density $U[\max(F_{min},\ F_i^{*\prime} - e_{F^*}), \min(1,\ F_i^{*\prime} + e_{F^*})]$.

*Updating adult non-null missing data rates (**β***)*

We updated the rate of adult missing data (not due to null alleles) at all loci at every MH iteration, one locus at a time. For each locus $l$, we propose $\beta_l^{*\prime} = u$, where $u \sim U\,[\max(0,\ \beta_l^{*} - e_{\beta^*}), \min(1,\ \beta_l^{*} + e_{\beta^*})]$ and $e_{\beta^*}$ is some incremental value. Denote $\boldsymbol{\beta}^{*\prime}$ the vector $\boldsymbol{\beta}^{*}$ with element $\beta_l^{*}$ replaced with $\beta_l^{*\prime}$, the move is then accepted with probability

$$\alpha(\boldsymbol{\beta}^{*\prime}|\,\boldsymbol{\beta}^{*}) = \min\left(1, \frac{\Pr(\mathbf{Y}\,|\,\mathbf{p}, \mathbf{F}^{*}, \boldsymbol{\beta}^{*\prime}, \boldsymbol{\mu}) f(\boldsymbol{\beta}^{*\prime}) q(\boldsymbol{\beta}^{*}, \boldsymbol{\beta}^{*\prime})}{\Pr(\mathbf{Y}\,|\,\mathbf{p}, \mathbf{F}^{*}, \boldsymbol{\beta}^{*}, \boldsymbol{\mu}) f(\boldsymbol{\beta}^{*}) q(\boldsymbol{\beta}^{*\prime}, \boldsymbol{\beta}^{*})}\right),$$

where $q(\boldsymbol{\beta}^{*}, \boldsymbol{\beta}^{*\prime})$ is the uniform density described above and $q(\boldsymbol{\beta}^{*\prime}, \boldsymbol{\beta}^{*})$ is the reverse move uniform density $U[\max(0,\ \beta_l^{*\prime} - e_{\beta^*}), \min(1,\ \beta_l^{*\prime} + e_{\beta^*})]$.

## References

Faubet, P. & Gaggiotti, O.E. (2008) A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics*, **178**, 1491–1504.

Hastings, W.K. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal Chemical Physics*, **21**, 1087–1092.

**Appendix S4** *Monte Carlo analysis of method performance*

Each independent replicate of the stochastic simulations used to test the model performance involved the six steps described below, given the assumed number of external populations $I$, offspring ($D$) and adult ($\boldsymbol{Q}$) sample sizes, population differentiation level $F_{ST}$, proportions of offspring from different population origins $\mathbf{m}$, population allelic frequencies before dispersal $\mathbf{p}$, population inbreeding coefficients $\mathbf{F}$ and $\mathbf{F}^*$, non-null missing data rates $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, and typing error rates $\boldsymbol{\mu}$. In what follows, we distinguish between *actual* alleles and *observed* alleles (as determined by genotyping errors and the unobservable nature of null alleles), and we use $r$ to denote a uniform random variate on [0, 1].

**(1)** Parametric frequencies for (*actual*) adult alleles at the $l$-th locus and $i$-th population, $\mathbf{p}_{il} = \{p_{il-}, p_{il1}, p_{il2}, \ldots, p_{ilk_l}\}$, corresponding to a level of population differentiation $F_{\mathrm{ST}}$ were obtained by sampling from a Dirichlet distribution (Wright 1931; Faubet et al. 2007)

$$f(p_{il-}, p_{il1}, p_{il2}, \ldots, p_{ilk_l}) = \Gamma(\lambda)\prod_{h=1}^{k_l}\frac{p_{ilh}^{\lambda q_{lh}-1}}{\Gamma(\lambda q_{lh})}$$

with $\lambda = 1/F_{\mathrm{ST}} - 1$, and $q_{lh}$ being the global frequency of the $h$-th allele of the $l$-th locus across populations, assumed to equal $1/(k_l+1)$ (i.e., equifrequent alleles). The frequency of null alleles is denoted by $p_{il-}$, and there are $k_l$ non-null alleles. After obtaining parametric allelic frequencies for all populations at locus $l$, including the recipient population, global $F_{\mathrm{ST}}$ was calculated and, if not within 10% of the target $F_{\mathrm{ST}}$ value, new parametric frequencies were generated until this condition was satisfied. This step was repeated for each of the $l$ loci.

**(2)** The number of offspring from each population origin in the simulated offspring sample was drawn from a multinomial distribution with $2I+1$ classes with probabilities $\{m_{p1}, m_{p2}, \ldots, m_{pI}, m_{s1}, m_{s2}, \ldots, m_{sI}, m_0\}$ and $D$ trials.

**(3)** For every locus $l$ of every offspring in the simulated sample born to parents in population $i$ and $j$, draw a new $r$ and generate the two *actual* homologous alleles from either: (a) two independent draws from two multinomial distributions with respective class probabilities $\mathbf{p}_{il}$ and $\mathbf{p}_{jl}$, if $i \neq j$; (b) two independent draws from a multinomial distribution with class probabilities $\mathbf{p}_{il}$, if $i = j$ and $r > F_i$; or (c) a single draw from a multinomial distribution with class probabilities $\mathbf{p}_{il}$ used to produce two identical-by-descent alleles, if $i = j$ and $r < F_i$.

**(4)** For every locus $l$ of each of the $Q_i$ simulated sampled adults of population $i$, draw a new $r$ and generate two *actual* homologous alleles from either: (a) two independent draws from a multinomial with class probabilities $\mathbf{p}_{il}$, if $r > F_i^*$, or (b) a single draw from a multinomial distribution with class probabilities $\mathbf{p}_{il}$, if $r < F_i^*$.

**(5)** For every *actual* non-null allele at locus $l$ of every simulated genotype, draw a new $r$ and set the corresponding *observed* allele at the actual allele if $r > \mu_l$, or at any of the other $k_l - 1$ non-null alleles if $r < \mu_l$. Homozygotes for a null allele are *observed* as missing data, while heterozygotes carrying a null allele are observed as homozygous for the non-null *observed* allele.

**(6)** Given the simulated adult and offspring *observed* genotypes, we estimated posterior distributions of parameters, with corresponding means and 95% credible intervals (CI), using equation 6 and the Metropolis-Hastings (MH) algorithm described in Appendix S3. Pilot runs were used to tune up the proposal distributions to obtain acceptance rates between 40–50%, after which we run the MH algorithm for 260,000 iterations, discarding the first 10,000 as burn-in and thinning the remaining ones to every 25$^{th}$, yielding a final sample of 1,000 observations.

For each combination of assumed parameter values, the six simulation steps were repeated to generate $N = 250$ independent realizations of the process and their associated $\hat{\boldsymbol{\Theta}}$ values, used to calculate expected estimation errors by comparing against $\boldsymbol{\Theta}$. As an example, the expected bias, expected root mean square error (*RMSE*), and expected CI non-coverage rate (NCR) of $\hat{\mathbf{m}}_p$ were computed using the following equations (analogously for other parameters):

$$Bias(\hat{\mathbf{m}}_p) = \frac{1}{N \cdot I} \sum_{q=1}^{N} \sum_{i=1}^{I} \left( \hat{m}_{pi}^{k} - m_{pi} \right)$$

$$RMSE(\hat{\mathbf{m}}_p) = \left[ \frac{1}{N \cdot I} \sum_{q=1}^{N} \sum_{i=1}^{I} \left( \hat{m}_{pi}^{k} - m_{pi} \right)^2 \right]^{0.5}$$

$$NCR(\hat{\mathbf{m}}_p) = \frac{1}{N \cdot I} \sum_{q=1}^{N} \sum_{i=1}^{I} A\left[ CI(\hat{m}_{pi}^{k}), m_{pi} \right]$$

where $\hat{m}_{pi}^{k}$ is the estimated pollen migration rate from the *i*-th population obtained for the *k*-th replicate data set, *NCR* is the proportion of times that the 95% CI does not contain the assumed value, and *A* is an indicator function taking the value 1 if the CI of $\hat{m}_{pi}^{k}$ contains $m_{pi}$, or zero otherwise.

## References

Faubet, P., Waples, R.S. & Gaggiotti, O.E. (2007) Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular Ecology*, **16**, 1149-1166.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

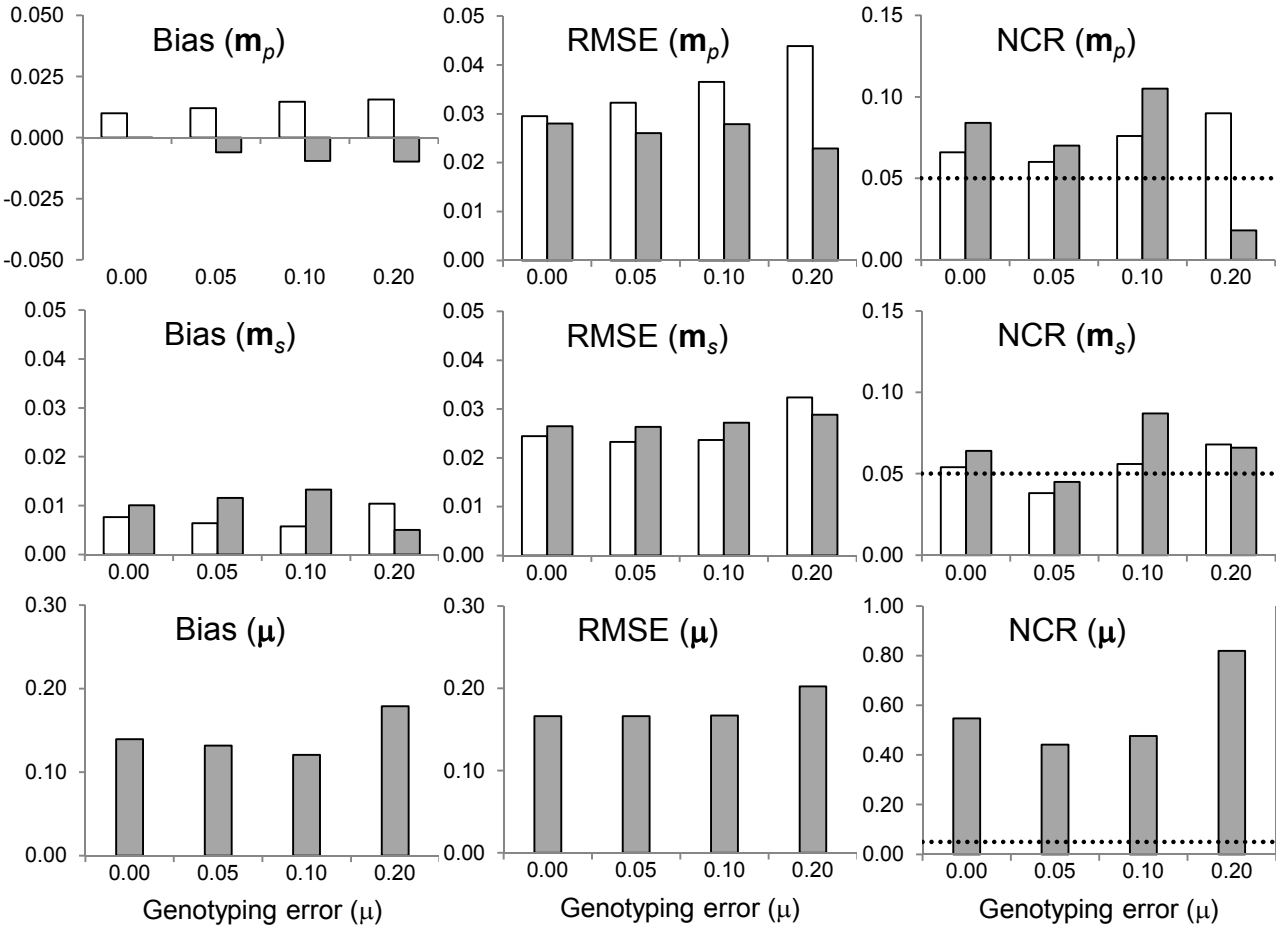


**Fig. S1**. Effect of genotyping error rate (μ) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained with $L$ = 20 loci (all with equal μ) and $k_l$ = 5 alleles/locus, by either ignoring (white bars) or jointly estimating (grey bars) μ. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I$ = 2 external populations, $m_{pi} = m_{si}$ = 0.05, $F_{ST}$ = 0.1, sample sizes of $D$ = 100 offspring and $Q_i$ = 100 adults/population.

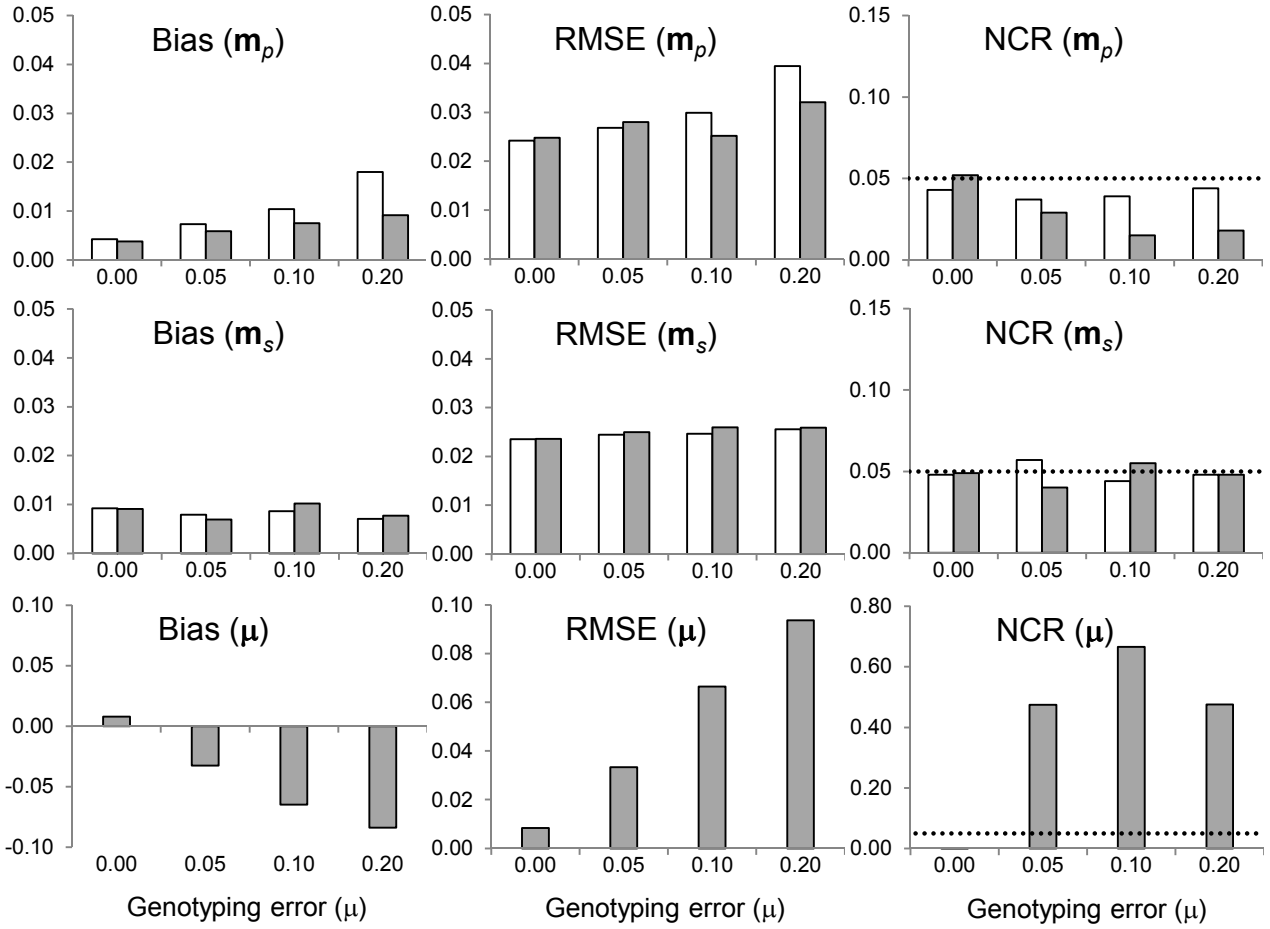


**Figure S2.** Effect of genotyping error rate (μ) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate estimates obtained with $L$ = 5 loci (all with equal μ) and $k_l$ = 20 alleles/locus, by either ignoring (white bars) or jointly estimating (grey bars) μ. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I$ = 2 external populations, $m_{pi} = m_{si}$ = 0.05, $F_{ST}$ = 0.1, sample sizes of $D$ = 100 offspring and $Q_i$ = 100 adults/population.

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

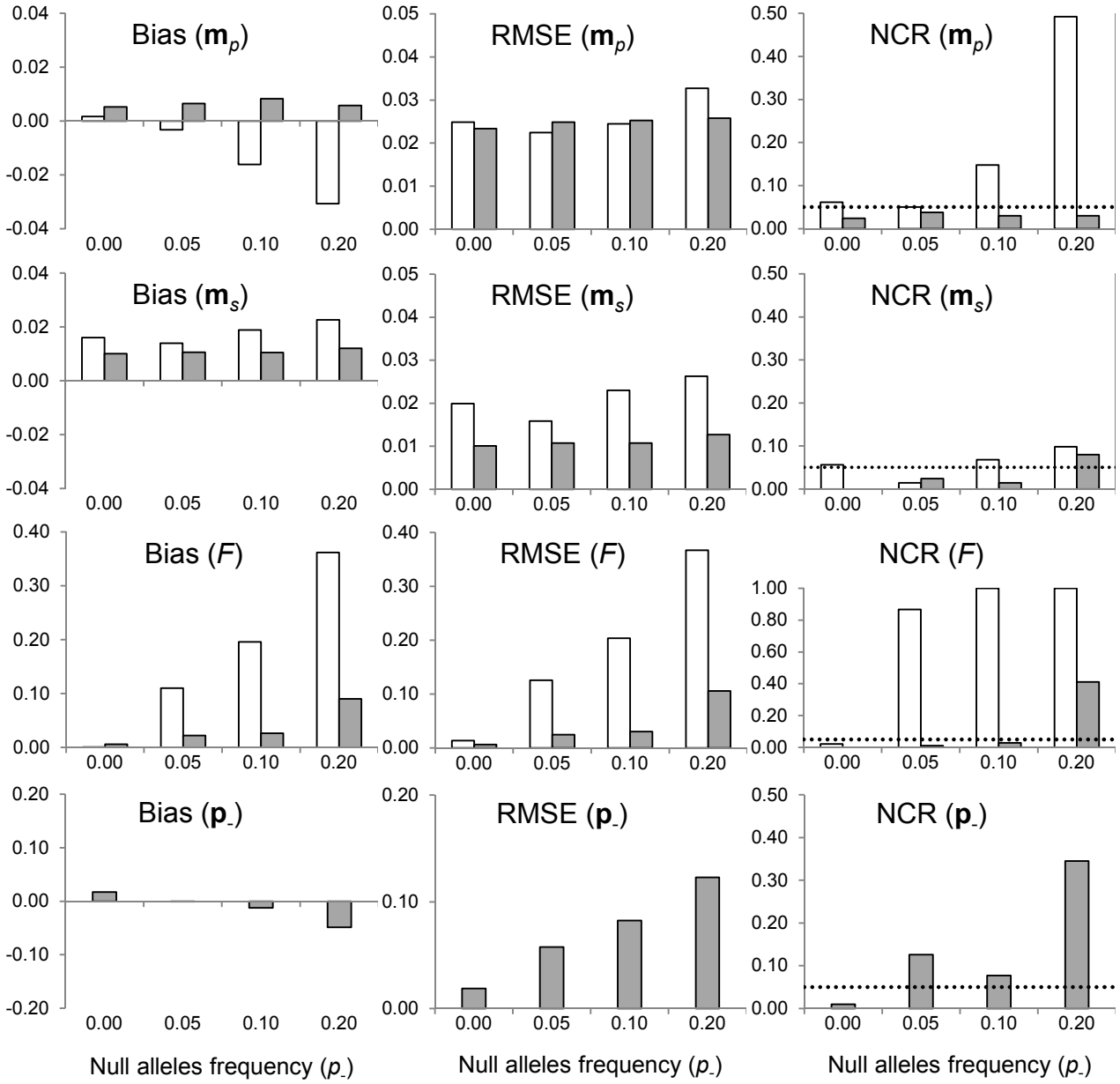


**Figure S3**. Effect of null alleles frequency ($p_-$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and population inbreeding ($F$) estimates obtained when $m_{si} = 0$ and $m_{pi} = 0.05$, by either ignoring (white bars) or jointly estimating (grey bars) $p_-$. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $F_{ST} = 0.1$, $F = 0$, $L = 5$ (for $p_- = 0.05$ scenario) or 10 (for $p_- \neq 0.05$ scenarios) loci, $k_l = 19$ (for $p_- = 0.05$), 10 (for $p_- = 0$), 9 (for $p_- = 0.1$), or 8 (for $p_- = 0.2$) non-null alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

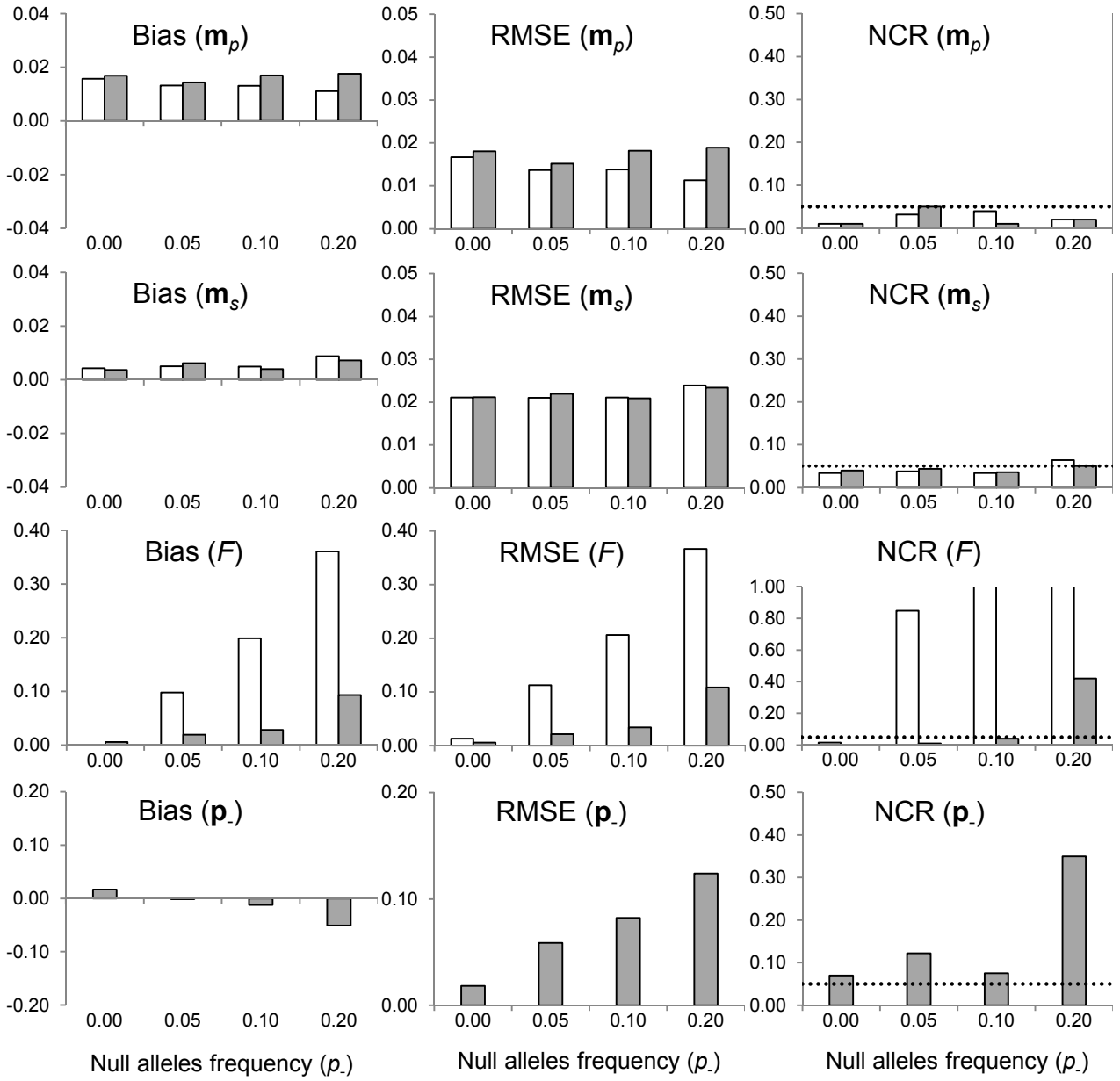


**Figure S4**. Effect of null alleles frequency ($p_-$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and population inbreeding ($F$) estimates obtained when $m_{si} = 0.05$ and $m_{pi} = 0$, by either ignoring (white bars) or jointly estimating (grey bars) $p_-$. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $F_{ST} = 0.1$, $F = 0$, $L = 5$ (for $p_- = 0.05$ scenario) or 10 (for $p_- \neq 0.05$ scenarios) loci, $k_l = 19$ (for $p_- = 0.05$), 10 (for $p_- = 0$), 9 (for $p_- = 0.1$), or 8 (for $p_- = 0.2$) non-null alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

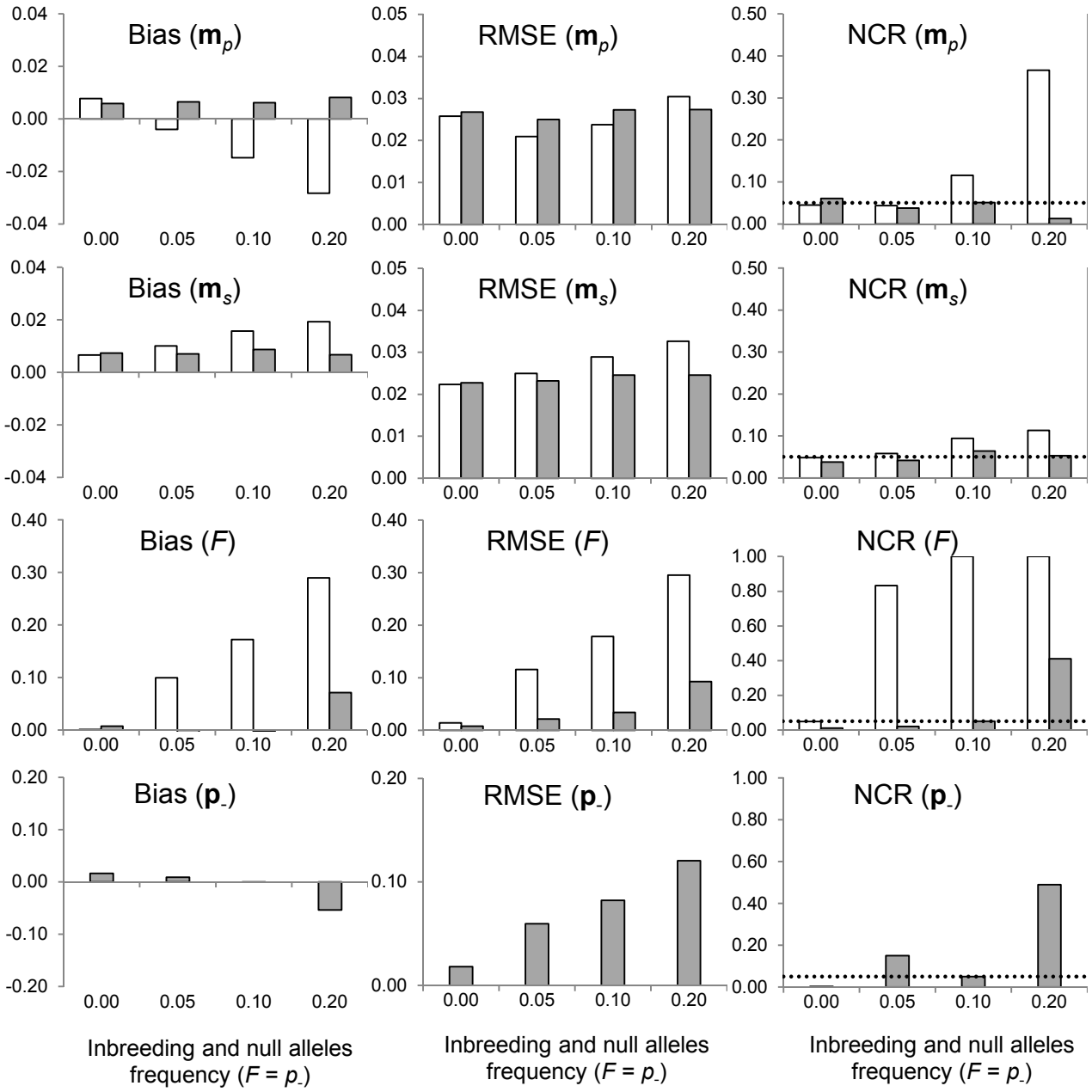


**Figure S5.** Effect of null alleles frequency ($p_-$) and population inbreeding ($F$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and $F$ estimates obtained by either ignoring (white bars) or jointly estimating (grey bars) $p_-$. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $m_{pi} = m_{si} = 0.05$, $F_{ST} = 0.1$, $L = 5$ (for $p_- = 0.05$ scenario) or 10 (for $p_- \neq 0.05$ scenarios) loci, $k_l = 19$ (for $p_- = 0.05$), 10 (for $p_- = 0$), 9 (for $p_- = 0.1$), or 8 (for $p_- = 0.2$) non-null alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.

**SUPPORTING INFORMATION** for *"Estimating contemporary migration rates: effect and joint inference of inbreeding, null alleles and mistyping"* by Juan J. Robledo-Arnuncio & Oscar Gaggiotti

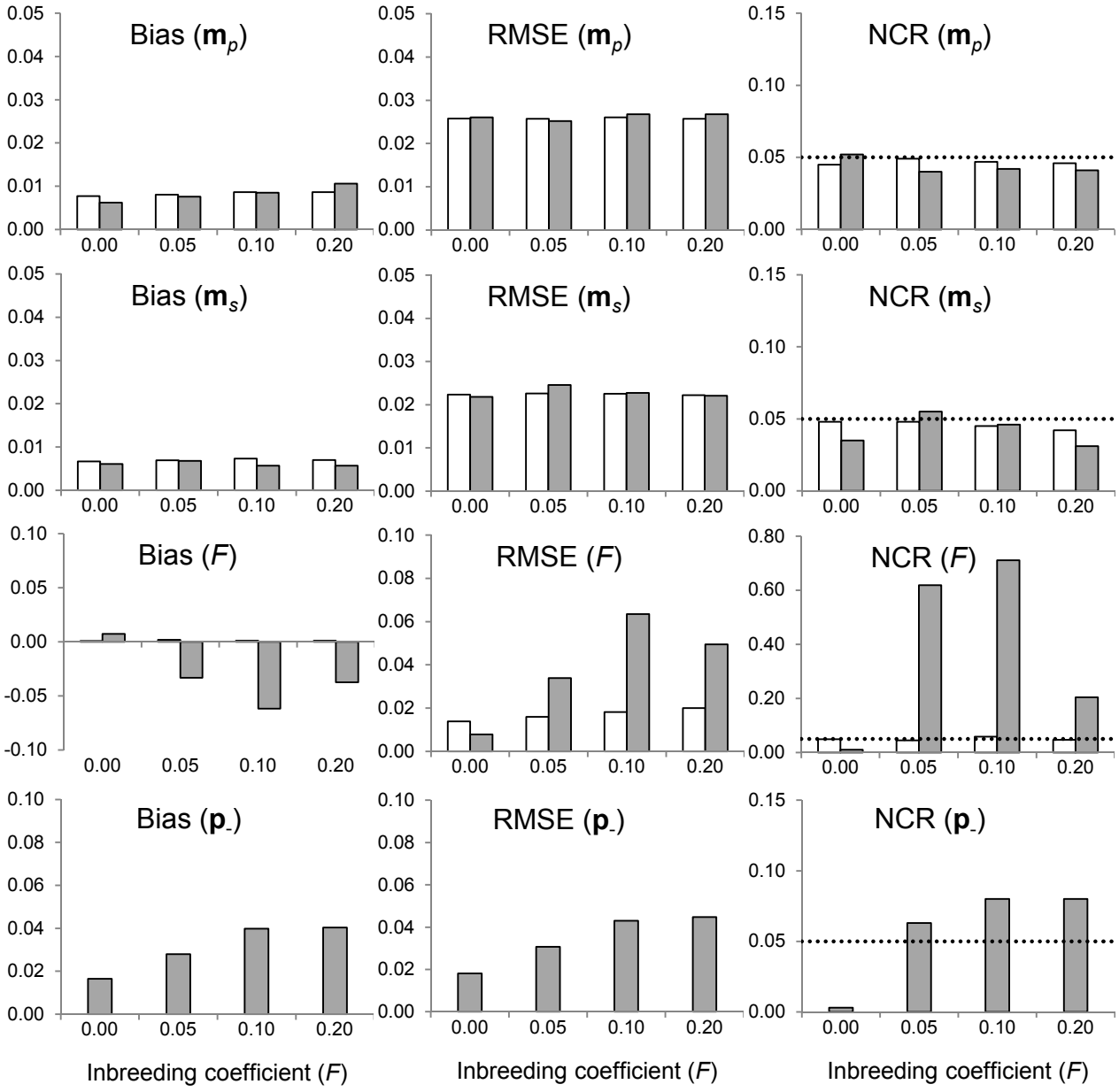


**Figure S6**. Effect of population inbreeding ($F$) on pollen ($\mathbf{m}_p$) and seed ($\mathbf{m}_s$) immigration rate and $F$ and $p_-$ (null alleles frequency) estimates obtained by either ignoring (white bars) or jointly estimating (grey bars) $p_-$. RMSE is the root mean square error and NCR the non-coverage rate of 95% credible intervals (the dotted line shows the nominal 5% value). Based on 250 Monte Carlo replicates per scenario, assuming: $I = 2$ external populations, $m_{pi} = m_{si} = 0.05$, $F_{ST} = 0.1$, $L = 10$ loci (all with $p_- = 0$), $k_l = 10$ non-null alleles/locus, sample sizes of $D = 100$ offspring and $Q_i = 100$ adults/population.